International Journal on Advances in Intelligent Systems

2012 vol. 5 nr. 3&4

The *International Journal on Advances in Intelligent Systems* is Published by IARIA.
ISSN: 1942-2679
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Intelligent Systems
Volume 5, Number 3 & 4, 2012

Julita Bermejo-Alonso, Universidad Politécnica de Madrid, Spain
Aurelio Bermúdez Marín, Universidad de Castilla-La Mancha, Spain
Lasse Berntzen, Vestfold University College - Tønsberg, Norway
Michela Bertolotto, University College Dublin, Ireland
Ateet Bhalla, Oriental Institute of Science & Technology, Bhopal, India
Freimut Bodendorf, Universität Erlangen-Nürnberg, Germany
Karsten Böhm, FH Kufstein Tirol - University of Applied Sciences, Austria
Pierre Borne, Ecole Centrale de Lille, France
Marko Bošković, Research Studios, Austria
Christos Bouras, University of Patras, Greece
Anne Boyer, LORIA - Nancy Université / KIWI Research team, France
Stainam Brandao, COPPE/Federal University of Rio de Janeiro, Brazil
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Vít Bršlica, University of Defence - Brno, Czech Republic
Dumitru Burdescu, University of Craiova, Romania
Diletta Romana Cacciagrano, University of Camerino, Italy
Kenneth P. Camilleri, University of Malta - Msida, Malta
Paolo Campegiani, University of Rome Tor Vergata , Italy
Marcelino Campos Oliveira Silva, Chemtech - A Siemens Business / Federal University of Rio de Janeiro, Brazil
Ozgu Can, Ege University, Turkey
José Manuel Cantera Fonseca, Telefónica Investigación y Desarrollo (R&D), Spain
Juan-Vicente Capella-Hernández, Universitat Politècnica de València, Spain
Bogdan Alexandru Caprarescu, West University of Timisoara, Romania
Miriam A. M. Capretz, The University of Western Ontario, Canada
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Davide Carboni, CRS4 Research Center - Sardinia, Italy
Mari Carmen Domingo, Barcelona Tech University, Spain
Luis Carriço, University of Lisbon, Portugal
Rafael Casado Gonzalez, Universidad de Castilla - La Mancha, Spain
Michelangelo Ceci, University of Bari, Italy
Fernando Cerdan, Polytechnic University of Cartagena, Spain
Alexandra Suzana Cernian, University "Politehnica" of Bucharest, Romania
Carlos Cetina, Technical Universidad San Jorge, Spain
Sukalpa Chanda, Gjøvik University College, Norway
David Chen, University Bordeaux 1, France
Luke Chen, University of Ulster @ Jordanstown, UK
Ping Chen, University of Houston-Downtown, USA
Kong Cheng, Telcordia Research, USA
Po-Hsun Cheng, National Kaohsiung Normal University, Taiwan
Dickson Chiu, Dickson Computer Systems, Hong Kong
Sunil Choenni, Research & Documentation Centre, Ministry of Security and Justice / Rotterdam University of Applied Sciences, The Netherlands
Ryszard S. Choras, University of Technology & Life Sciences, Poland
Smitashree Choudhury, Knowledge Media Institute, The UK Open University, UK
William Cheng-Chung Chu, Tunghai University, Taiwan
Christophe Claramunt, Naval Academy Research Institute, France

Cesar A. Collazos, Universidad del Cauca, Colombia
Phan Cong-Vinh, NTT University, Vietnam
Christophe Cruz, University of Bourgogne, France
Beata Czarnacka-Chrobot, Warsaw School of Economics, Department of Business Informatics, Poland
Claudia d'Amato, University of Bari, Italy
Sérgio Roberto P. da Silva, Universidade Estadual de Maringá - Paraná, Brazil
Mirela Danubianu, "Stefan cel Mare" University of Suceava, Romania
Dragos Datcu, Netherlands Defense Academy / Delft University of Technology , The Netherlands
Antonio De Nicola, ENEA, Italy
Claudio de Castro Monteiro, Federal Institute of Education, Science and Technology of Tocantins, Brazil
Noel De Palma, Joseph Fourier University, France
Jan Dedek, Charles University in Prague, Czech Republic
Zhi-Hong Deng, Peking University, China
Stojan Denic, Toshiba Research Europe Limited, UK
Vivek S. Deshpande, MIT College of Engineering - Pune, India
Sotirios Ch. Diamantas, Pusan National University, South Korea
Leandro Dias da Silva, Universidade Federal de Alagoas, Brazil
Jerome Dinet, Univeristé Paul Verlaine - Metz, France
Jianguo Ding, University of Luxembourg, Luxembourg
Yulin Ding, Defence Science & Technology Organisation Edinburgh, Australia
Alexiei Dingli, University of Malta, Malta
Mihaela Dinsoreanu, Technical University of Cluj-Napoca, Romania
Ioanna Dionysiou, University of Nicosia, Cyprus
Roland Dodd, CQUniversity, Australia
Nima Dokoohaki, Royal Institute of Technology (KTH)-Kista, Sweden
Suzana Dragicevic, Simon Fraser University- Burnaby, Canada
Mauro Dragone, University College Dublin (UCD), Ireland
Marek J. Druzdzel, University of Pittsburgh, USA
Carlos Duarte, University of Lisbon, Portugal
Raimund K. Ege, Northern Illinois University, USA
Jorge Ejarque, Barcelona Supercomputing Center, Spain
Larbi Esmahi, Athabasca University, Canada
Simon G. Fabri, University of Malta, Malta
Umar Farooq, Amazon.com, USA
Mehdi Farshbaf-Sahih-Sorkhabi, Azad University - Tehran / Fanavaran co., Tehran, Iran
Anna Fensel, Semantic Technology Institute (STI) Innsbruck and FTW Forschungszentrum Telekommunikation Wien, Austria
Stenio Fernandes, Federal University of Pernambuco (CIn/UFPE), Brazil
Oscar Ferrandez Escamez, University of Utah, USA
Florin Filip, Romanian Academy, Romania
Agata Filipowska, Poznan University of Economics, Poland
Ziny Flikop, Scientist, USA
Adina Magda Florea, University "Politehnica" of Bucharest, Romania
Francesco Fontanella, University of Cassino and Southern Lazio, Italy
Panagiotis Fotaris, University of Macedonia, Greece
Enrico Francesconi, ITTIG - CNR / Institute of Legal Information Theory and Techniques / Italian National Research

Council, Italy
Rita Francese, Università di Salerno - Fisciano, Italy
Bernhard Freudenthaler, Software Competence Center Hagenberg GmbH, Austria
Sören Frey, University of Kiel, Germany
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Somchart Fugkeaw, Thai Digital ID Co., Ltd., Thailand
Naoki Fukuta, Shizuoka University, Japan
Mathias Funk, Eindhoven University of Technology, The Netherlands
Adam M. Gadomski**,** Università degli Studi di Roma La Sapienza, Italy
Alex Galis, University College London (UCL), UK
Crescenzio Gallo, Department of Clinical and Experimental Medicine - University of Foggia, Italy
Matjaz Gams, Jozef Stefan Institute-Ljubljana, Slovenia
Raúl García Castro, Universidad Politécnica de Madrid, Spain
Fabio Gasparetti, Roma Tre University - Artificial Intelligence Lab, Italy
Joseph A. Giampapa, Carnegie Mellon University, USA
George Giannakopoulos, NCSR Demokritos, Greece
David Gil, University of Alicante, Spain
Harald Gjermundrod, University of Nicosia, Cyprus
Angelantonio Gnazzo, Telecom Italia - Torino, Italy
Luis Gomes, Universidade Nova Lisboa, Portugal
Nan-Wei Gong, MIT Media Laboratory, USA
Francisco Alejandro Gonzale-Horta, National Institute for Astrophysics, Optics, and Electronics (INAOE), Mexico
Sotirios K. Goudos, Aristotle University of Thessaloniki, Greece
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Gregor Grambow, University of Ulm, Germany
Fabio Grandi, University of Bologna, Italy
Andrina Granić, University of Split, Croatia
Carmine Gravino, Università degli Studi di Salerno, Italy
Dominic Greenwood, Whitestein Technologies, Switzerland
Michael Grottke, University of Erlangen-Nuremberg, Germany
Vic Grout, Glyndŵr University, UK
Maik Günther, Stadtwerke München GmbH, Germany
Francesco Guerra, University of Modena and Reggio Emilia, Italy
Alessio Gugliotta, Innova SPA, Italy
Richard Gunstone, Bournemouth University, UK
Fikret Gurgen, Bogazici University, Turkey
Ivan Habernal, University of West Bohemia, Czech Republic
Maki Habib, The American University in Cairo, Egypt
Till Halbach Røssvoll, Norwegian Computing Center, Norway
Jameleddine Hassine, King Fahd University of Petroleum & Mineral (KFUPM), Saudi Arabia
Ourania Hatzi, Harokopio University of Athens, Greece
Yulan He, The Open University, UK
Kari Heikkinen, Lappeenranta University of Technology, Finland
Cory Henson, Wright State University / Kno.e.sis Center, USA
Arthur Herzog, Technische Universität Darmstadt, Germany
Rattikorn Hewett, Whitacre College of Engineering, Texas Tech University, USA

Celso Massaki Hirata, Instituto Tecnológico de Aeronáutica - São José dos Campos, Brazil
Jochen Hirth, University of Kaiserslautern, Germany
Bernhard Hollunder, Hochschule Furtwangen University, Germany
Thomas Holz, University College Dublin, Ireland
Władysław Homenda, Warsaw University of Technology, Poland
Carolina Howard Felicíssimo, Schlumberger Brazil Research and Geoengineering Center, Brazil
Jingwei Huang , University of Illinois at Urbana-Champaign, USA
Weidong (Tony) Huang, CSIRO ICT Centre, Australia
Xiaodi Huang, Charles Sturt University - Albury, Australia
Eduardo Huedo, Universidad Complutense de Madrid, Spain
Marc-Philippe Huget, University of Savoie, France
Chi Hung, Tsinghua University, China
Chih-Cheng Hung, Southern Polytechnic State University - Marietta, USA
Edward Hung, Hong Kong Polytechnic University, Hong Kong
Muhammad Iftikhar, Universiti Malaysia Sabah (UMS), Malaysia
Prateek Jain, Ohio Center of Excellence in Knowledge-enabled Computing, Kno.e.sis, USA
Wassim Jaziri, Miracl Laboratory, ISIM Sfax, Tunisia
Hoyoung Jeung, SAP Research Brisbane, Australia
Yiming Ji, University of South Carolina Beaufort, USA
Jinlei Jiang, Department of Computer Science and Technology, Tsinghua University, China
Weirong Jiang, Juniper Networks Inc., USA
Hanmin Jung, Korea Institute of Science & Technology Information, Korea
Ilya S. Kabak, "Stankin" Moscow State Technological University, Russia
Eleanna Kafeza, Athens University of Economics and Business, Greece
Hermann Kaindl, Vienna University of Technology, Austria
Ahmed Kamel, Concordia College, Moorhead, Minnesota, USA
Faouzi Kamoun, University of Dubai, UAE
Rajkumar Kannan, Bishop Heber College(Autonomous), India
Teemu Kanstrén, VTT, Finland
Fazal Wahab Karam, Norwegian University of Science and Technology (NTNU), Norway
Dimitrios A. Karras, Chalkis Institute of Technology, Hellas
Koji Kashihara, The University of Tokushima, Japan
Nittaya Kerdprasop, Suranaree University of Technology, Thailand
Katia Kermanidis, Ionian University, Greece
Serge Kernbach, University of Stuttgart, Germany
Nhien An Le Khac, University College Dublin, Ireland
Malik Jahan Khan, Lahore University of Management Sciences (LUMS), Lahore, Pakistan
Reinhard Klemm, Avaya Labs Research, USA
Ah-Lian Kor, Leeds Metropolitan University, UK
Arne Koschel, Applied University of Sciences and Arts, Hannover, Germany
George Kousiouris, NTUA, Greece
Philipp Kremer, German Aerospace Center (DLR), Germany
Dalia Kriksciuniene, Vilnius University, Lithuania
Dariusz Król, AGH University of Science and Technology, ACC Cyfronet AGH, Poland
Roland Kübert, Höchstleistungsrechenzentrum Stuttgart, Germany
Markus Kunde, German Aerospace Center, Germany

Dharmender Singh Kushwaha, Motilal Nehru National Institute of Technology, India
Andrew Kusiak, The University of Iowa, USA
Dimosthenis Kyriazis, National Technical University of Athens, Greece
Vitaveska Lanfranchi, Research Fellow, OAK Group, University of Sheffield, UK
Mikel Larrea, University of the Basque Country UPV/EHU, Spain
Angelos Lazaris, University of Southern California, USA
Philippe Le Parc, University of Brest, France
Gyu Myoung Lee, Institut Telecom, Telecom SudParis, France
Kyu-Chul Lee, Chungnam National University, South Korea
Tracey Kah Mein Lee, Singapore Polytechnic, Republic of Singapore
Daniel Lemire, LICEF Research Center, Canada
Haim Levkowitz, University of Massachusetts Lowell, USA
Kuan-Ching Li, Providence University, Taiwan
Tsai-Yen Li, National Chengchi University, Taiwan
Yangmin Li, University of Macau, Macao SAR
Jian Liang, Nimbus Centre, Cork Institute of Technology, Ireland
Haibin Liu, China Aerospace Science and Technology Corporation, China
Lu Liu, University of Derby, UK
Qing Liu, The Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Shih-Hsi "Alex" Liu, California State University - Fresno, USA
Xiaoqing (Frank) Liu, Missouri University of Science and Technology, USA
David Lizcano, Universidad a Distancia de Madrid, Spain
Henrique Lopes Cardoso, LIACC / Faculty of Engineering, University of Porto, Portugal
Wassef Louati, University of Monastir, Tunisia
Sandra Lovrencic, University of Zagreb, Croatia
Jun Luo, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China
Prabhat K. Mahanti, University of New Brunswick, Canada
Jacek Mandziuk, Warsaw University of Technology, Poland
Herwig Mannaert, University of Antwerp, Belgium
Yannis Manolopoulos, Aristotle University of Thessaloniki, Greece
Antonio Maria Rinaldi, Università di Napoli Federico II, Italy
Ali Masoudi-Nejad, University of Tehran, Iran
Constandinos Mavromoustakis, University of Nicosia, Cyprus
Gerrit Meixner, German Research Center for Artificial Intelligence (DFKI) / Innovative Factory Systems (IFS) / Center for Human-Machine-Interaction (ZMMI), Germany
Zulfiqar Ali Memon, Sukkur Institute of Business Administration, Pakistan
Andreas Merentitis, AGT Group (R&D) GmbH, Germany
Jose Merseguer, Universidad de Zaragoza, Spain
Frederic Migeon, IRIT/Toulouse University, France
Harald Milchrahm, Technical University Graz, Institute for Software Technology, Austria
Fatma Mili, Oakland University, USA
Les Miller, Iowa State University, USA
Marius Minea, University POLITEHNICA of Bucharest, Romania
Yasser F. O. Mohammad, Assiut University, Egypt
Shahab Mokarizadeh, Royal Institute of Technology (KTH) - Stockholm, Sweden
Martin Molhanec, Czech Technical University in Prague, Czech Republic

Dorothy Monekosso, University of Ulster at Jordanstown, UK
Charalampos Moschopoulos, KU Leuven, Belgium
Mary Luz Mouronte López, Ericsson S.A., Spain
Henning Müller, University of Applied Sciences Western Switzerland - Sierre (HES SO), Switzerland
Susana Munoz Hernández, Universidad Politécnica de Madrid, Spain
Adrian Muscat, University of Malta, Malta
Peter Mutschke, GESIS - Leibniz Institute for the Social Sciences - Bonn, Germany
Bela Mutschler, Hochschule Ravensburg-Weingarten, Germany
Deok Hee Nam, Wilberforce University, USA
Fazel Naghdy, University of Wollongong, Australia
Joan Navarro, Research Group in Distributed Systems (La Salle - Ramon Llull University), Spain
Saša Nešić, University of Lugano, Switzerland
Rui Neves Madeira, Instituto Politécnico de Setúbal / Universidade Nova de Lisboa, Portugal
Toàn Nguyên, INRIA Grenobel Rhone-Alpes/ Montbonnot, France
Andrzej Niesler, Institute of Business Informatics, Wroclaw University of Economics, Poland
Michael P. Oakes, University of Sunderland, UK
John O'Donovan, University of California - Santa Barbara, USA
Kouzou Ohara, Aoyama Gakuin University, Japan
Jonice Oliveira, Universidade Federal do Rio de Janeiro, Brazil
Ian Oliver, Nokia Location & Commerce, Finland / University of Brighton, UK
Michael Adeyeye Oluwasegun, University of Cape Town, South Africa
Sigeru Omatu, Osaka Institute of Technology, Japan
Sascha Opletal, University of Stuttgart, Germany
Flavio Oquendo, European University of Brittany/IRISA-UBS, France
Fakri Othman, Cardiff Metropolitan University, UK
Enn Õunapuu, Tallinn University of Technology, Estonia
Jeffrey Junfeng Pan, Facebook Inc., USA
Hervé Panetto, University of Lorraine, France
Malgorzata Pankowska, University of Economics, Poland
Harris Papadopoulos, Frederick University, Cyprus
Laura Papaleo, ICT Department - Province of Genoa & University of Genoa, Italy
Agis Papantoniou, National Technical University of Athens, Greece
Thanasis G. Papaioannou, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Andreas Papasalouros, University of the Aegean, Greece
Eric Paquet, National Research Council / University of Ottawa, Canada
Kunal Patel, Ingenuity Systems, USA
Carlos Pedrinaci, Knowledge Media Institute, The Open University, UK
Yoseba Penya, University of Deusto - DeustoTech (Basque Country), Spain
Cathryn Peoples, University of Ulster, UK
Asier Perallos, University of Deusto, Spain
Christian Percebois, Université Paul Sabatier - IRIT, France
Andrea Perego, European Commission, Joint Research Centre, Italy
Mark Perry, University of Western Ontario/Faculty of Law/ Faculty of Science - London, Canada
Willy Picard, Poznań University of Economics, Poland
Meikel Poess, Oracle, USA
Agostino Poggi, Università degli Studi di Parma, Italy

R. Ponnusamy, Madha Engineering College-Anna University, India
Dorin Popescu, University of Craiova, Romania
Stefan Poslad, Queen Mary University of London, UK
Wendy Powley, Queen's University, Canada
Radu-Emil Precup, "Politehnica" University of Timisoara, Romania
Jerzy Prekurat, Canadian Bank Note Co. Ltd., Canada
Didier Puzenat, Université des Antilles et de la Guyane, France
Sita Ramakrishnan, Monash University, Australia
Elmano Ramalho Cavalcanti, Federal University of Campina Grande, Brazil
Juwel Rana, Luleå University of Technology, Sweden
Martin Randles, School of Computing and Mathematical Sciences, Liverpool John Moores University, UK
Christoph Rasche, University of Paderborn, Germany
Ann Reddipogu, ManyWorlds UK Ltd, UK
Ramana Reddy, West Virginia University, USA
René Reiners, Fraunhofer FIT - Sankt Augustin, Germany
Paolo Remagnino, Kingston University - Surrey, UK
Sebastian Rieger, Karlsruher Institut für Technologie (KIT) / Steinbuch Centre for Computing (SCC), Germany
Andreas Riener, Johannes Kepler University Linz, Austria
Ivan Rodero, NSF Center for Autonomic Computing, Rutgers University - Piscataway, USA
Joel Rodrigues, Instituto de Telecomunicações / University of Beira Interior, Portugal
Alejandro Rodríguez González, University Carlos III of Madrid, Spain
Aitor Rodríguez-Alsina, University Autonoma of Barcelona (UAB), Spain
Paolo Romano, INESC-ID Lisbon, Portugal
Vicente-Arturo Romero-Zaldivar, Atos Origin SAE, Spain
Agostinho Rosa, Instituto de Sistemas e Robótica, Portugal
José Rouillard, University of Lille, France
Paweł Różycki, University of Information Technology and Management (UITM) in Rzeszów, Poland
Igor Ruiz-Agundez, DeustoTech, University of Deusto, Spain
Michele Ruta, Politecnico di Bari, Italy
Melike Sah, Trinity College Dublin, Ireland
Francesc Saigi Rubió, Universitat Oberta de Catalunya, Spain
Abdel-Badeeh M. Salem, Ain Shams University, Egypt
Yacine Sam, Université François-Rabelais Tours, France
Ismael Sanz, Universitat Jaume I, Spain
Ricardo Sanz, Universidad Politecnica de Madrid, Spain
Marcello Sarini, Università degli Studi Milano-Bicocca - Milano, Italy
Munehiko Sasajima, I.S.I.R., Osaka University, Japan
Minoru Sasaki, Ibaraki University, Japan
Hiroyuki Sato, University of Tokyo, Japan
Jürgen Sauer, Universität Oldenburg, Germany
Patrick Sayd, CEA List, France
Dominique Scapin, INRIA - Le Chesnay, France
Kenneth Scerri, University of Malta, Malta
Adriana Schiopoiu Burlea, University of Craiova, Romania
Rainer Schmidt, Austrian Institute of Technology, Austria
Bruno Schulze, National Laboratory for Scientific Computing - LNCC, Brazil

Wieland Schwinger, Johannes Kepler University Linz, Austria
Hans-Werner Sehring, T-Systems Multimedia Solutions GmbH, Germany
Paulo Jorge Sequeira Gonçalves, Polytechnic Institute of Castelo Branco, Portugal
Sandra Sendra Compte, Polytechnic University of Valencia, Spain
Kewei Sha, Oklahoma City University, USA
Hossein Sharif, University of Portsmouth, UK
Roman Y. Shtykh, Rakuten, Inc., Japan
Kwang Mong Sim, Gwangju Insitute of Science & Technology, South Korea
Robin JS Sloan, University of Abertay Dundee, UK
Vasco N. G. J. Soares, Instituto de Telecomunicações / University of Beira Interior / Polytechnic Institute of Castelo Branco, Portugal
Don Sofge, Naval Research Laboratory, USA
Christoph Sondermann-Woelke, Universitaet Paderborn, Germany
George Spanoudakis, City University London, UK
Vladimir Stantchev, SRH University Berlin, Germany
Claudius Stern, University of Paderborn, Germany
Mari Carmen Suárez-Figueroa, Universidad Politécnica de Madrid (UPM), Spain
Kåre Synnes, Luleå University of Technology, Sweden
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Yehia Taher, ERISS - Tilburg University, The Netherlands
Yutaka Takahashi, Senshu University, Japan
Azzelarabe Taleb-Bendiab, Liverpool John Moores University, UK
Dan Tamir, Texas State University, USA
Jinhui Tang, Nanjing University of Science and Technology, P.R. China
Yi Tang, Chinese Academy of Sciences, China
Said Tazi, LAAS-CNRS, Universite Toulouse 1, France
John Terzakis, Intel, USA
Sotirios Terzis, University of Strathclyde, UK
Vagan Terziyan, University of Jyvaskyla, Finland
Michael Tighe, University of Western Ontario, Canada
Ioan Toma, STI Innsbruck/University Innsbruck, Austria
Lucio Tommaso De Paolis, Department of Innovation Engineering - University of Salento, Italy
Davide Tosi, Università degli Studi dell'Insubria, Italy
Raquel Trillo Lado, University of Zaragoza, Spain
Tuan Anh Trinh, Budapest University of Technology and Economics, Hungary
Simon Tsang, Applied Communication Sciences, USA
Theodore Tsiligiridis, Agricultural University of Athens, Greece
Antonios Tsourdos, Cranfield University, UK
José Valente de Oliveira, University of Algarve, Portugal
Cristián Felipe Varas Schuda, NIC Chile Research Labs, Chile
Eugen Volk, University of Stuttgart, Germany
Mihaela Vranić, University of Zagreb, Croatia
Chieh-Yih Wan, Intel Labs, Intel Corporation, USA
Jue Wang, Washington University in St. Louis, USA
Shenghui Wang, OCLC Leiden, The Netherlands
Zhonglei Wang, Karlsruhe Institute of Technology (KIT), Germany

International Journal on Advances in Intelligent Systems
Volume 5, Numbers 3 & 4, 2012

CONTENTS

An Efficient Bi-Dimensional Indexing Scheme for Three-Dimensional Trajectories

Antonio d'Acierno
Institute of Food Science
National Research Council of Italy
Avellino, Italy
dacierno.a@isa.cnr.it

Marco Leone, Alessia Saggese,
Mario Vento
DIEII, University of Salerno
Fisciano (SA), Italy
{*mleone,asaggese,mvento*}*@unisa.it*

***Abstract*—The proliferation of devices able to monitor their position is favoring the accumulation of large amount of geographically referenced data, that can be profitably used in a lot of applications, ranging from traffic control and management to location-aware services. The strong interest in these applications has entailed a significant research effort in the last years, both toward the modeling of spatio-temporal databases and toward indexing strategies to efficiently process spatio-temporal queries. Recently, we presented an indexing scheme (based on a redundant storing strategy) able to index three-dimensional trajectories using widely available bidimensional indexes. In this paper we propose a method that, while avoids redundant storing of data, still uses well established bi-dimensional indexes. With respect to the previous work, the retrieving performance is improved by taking advantage both of a more efficient representation and of a trajectory segmentation stage, as experimental results show.**

Keywords-indexing; moving objects databases; spatial queries.

I. INTRODUCTION

The increasing number of mobile devices able to report their position in real-time with high accuracy has implied the collection of large amount of data, which can be profitably used in many realms, ranging from traffic control and management to location-aware services [1][2]. Moreover, the need for security in many public environments has also contributed to an exponential proliferation in the number of available cameras and, starting from the video streams acquired from these peripherals, objects' positions can be extracted by using available video analytic algorithms [3]. In this way, databases for the analysis and the validation of models related to different typologies of objects' movements and behaviors (pedestrians, cars, pets and so on) have gained great interest.

In order to store and efficiently retrieve the information extracted from this large amount of acquired data, in the last years a significant research effort has been made, both towards the modeling of spatio-temporal Moving Object Databases (MODs) and towards indexing strategies aimed at efficiently process spatio-temporal queries.

According to the widely adopted line segment model, it is straightforward to represent the movement of each object as a sequence of line segments, each represented by two sample positions at consecutive time instants. The trajectory associated to the object's motion is thus represented by a polyline in a three-dimensional space, the first two dimensions referring to the space and the third one to the time. An example is shown in Figure 1.

When handling with MODs, we typically aim at extracting, from the entire collection of stored data, only those trajectories possessing a given property: information retrieval is therefore achieved through processing the query submitted by the user.

Queries that are worth to be considered in spatio-temporal databases can be subdivided at least into three orthogonal categories. First, from a temporal perspective, *"find all the vehicles that will be in a given area in the next ten minutes"* is an instance of the so-called *future query*; in order to solve this query, models able to predict the future position of a moving object are needed. On the other hand, *past queries* handle with the historical positions of moving objects. Finally, *now queries* ask for the position of objects at the current time instant t^{now}; these queries can be considered as a special case of future queries if the last recorded position is lower than t^{now}; furthermore, they represent a special case (at least from the indexing strategy point of view) of now queries if the current position of objects has been recorded.

Another commonly accepted query taxonomy [4][5] is based on the following consideration: a trajectory is a very complex data structure, so implying that the time needed to extract it from the database strongly influences the performance of a generic retrieval system. For this reason, it is important to distinguish, for example, if we are only interested in the identifiers of the objects passing through a given area in a given time interval or if we are interested in the whole trajectory of these objects. The former query is commonly defined as *coordinate-based query*; a typical example is: *'find the number of pedestrians in Saint Peter's square in Rome between 9-12 am yesterday'*; the latter is commonly defined as *trajectory-based* query and, in turn, contains two different categories: *topological queries* and *navigational queries*. Topological queries make use of information about the scene for the extraction of an object's trajectory (*'when did vehicle X leave Plebiscito's square in Naples most recently?'*). Navigational queries need derived

information to answer questions like *'what is the current speed of vehicle X?'*; in this case, the needed information (like speed, heading, travel distance, etc.) is not typically stored directly and, therefore, a computational overhead is paid.

Furthermore, from a temporal point of view, a simple but useful distinction is made between time slice and time interval queries; a *time slice query* asks about a fixed time point while a *time interval query* considers a temporal interval. For the sake of clarity, a *time slice query*, in its coordinate-based form, can be exemplified as *'find all the objects that are in a given area at a given time instant* t_i*'* while its trajectory-based form is *'find the trajectories of each object that is in a given area at a given time instant* t_i*'*. When $t_i < t_{now}$, we are facing a *past query*; $t_i > t^{now}$ characterizes *future queries*, while $t_i = t^{now}$ is the case of *now queries*. A *time interval query*, in its coordinate-based form, can be exemplified as *'find all the objects that pass through a given area in the time interval* $[t_1, t_2]$*'*, while a trajectory-based example is *'find all the trajectories of objects that traverse a given area in the time interval* $[t_1, t_2]$*'*. Assuming $t_1 < t_2$, if $t_2 < t^{now}$ we are dealing with a *past query*, while $t_1 > t^{now}$ characterizes *future queries*. The case $t_1 < t^{now} < t_2$ can be easily assumed as composed by a past time interval query ($t \in [t_1, t^{now})$, a now time slice query ($t = t^{now}$) and a future time interval query ($t \in (t^{now}, t_2]$).

Many other interesting queries are reported in [6], [7], [8], [9] and in [10]. The large amount of queries that have been proposed, especially in the last years, reinforces the evidence that research is still ongoing in this field and, at our knowledge, efficient solutions are still being investigated.

In this context, we are interested in efficiently storing and querying moving objects' trajectories extracted from video cameras. Although efficient bidimensional indexing methods are usually available, several problems arise when data to be handled are three- or even four-dimensional, as it happens for the trajectory-based systems. To solve these problems, we recently proposed a method [1] able to redundantly project and analyze a collection of trajectories on bi-dimensional planes by using off-the-shelf solutions.

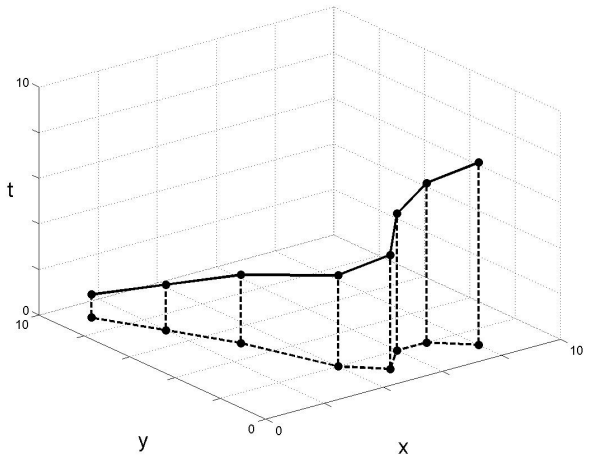


Figure 1. A spatio-temporal trajectory; *x* and *y* dimensions refer to position while the third dimension (*t*) refers to time.

Starting from our previous work, in this paper we propose an improved version of the system able to answer past (trajectory-based) time interval queries on a MOD; even still using bidimensional indexes, the proposed solution avoids the redundancy in the stored data and improves the whole performance, also thanks to a segmentation algorithm aimed at optimizing the use of the adopted indexes.

The paper is organized as follows: after a discussion about some of the papers related to both bi-dimensional and three-dimensional data indexing methods (Section II), in Section III we describe our previous proposal in [1] and its improvement, obtained by using a new indexing strategy that avoids redundancy in the data to be stored; in Section III we also introduce and describe the adopted segmentation algorithm. Experimental results are the concern of Section IV, while Section V concludes the paper outlining future directions.

II. Related Work

Indexing moving objects databases has been an active research area in the recent past and several solutions have been proposed. [6] and [11] survey many accessing strategies, proposed in the last two decades, which are able to index the past and the current position, as well as methods supporting queries related to the future.

According to [11] and [12], one of the most influential accessing methods in the area of spatial data management was proposed by Guttman. He suggested, in his pioneering paper [13], a structure named R-tree able to efficiently index bidimensional rectangular objects in VLSI (Very Large Scale Integration) design applications. The conceptual simplicity of an R-tree and its resemblance to widely adopted standard B-trees, allowed the developers to easily incorporate such a solution in spatial enabled DBMS [12] in order to support spatial query optimization and processing. R-trees hierarchically organize the geometric objects by representing them through *Minimum Bounding Rectangles* (MBRs, [14]), which are an expression of the object's maximum extents in its coordinate system; each internal node corresponds to the MBR that bounds its children while, as usual, a leaf contains pointers to the objects (see Figure 2). The insertion of a new object takes place by choosing, at each level, the node that involves the smallest expansion; when a split of the selected leaf node is needed, Guttman proposed three algorithms with different complexity to handle such a split, aiming at the minimization of the sum of the areas of resulting nodes. It is worth noting that, since an MBR can be included in many

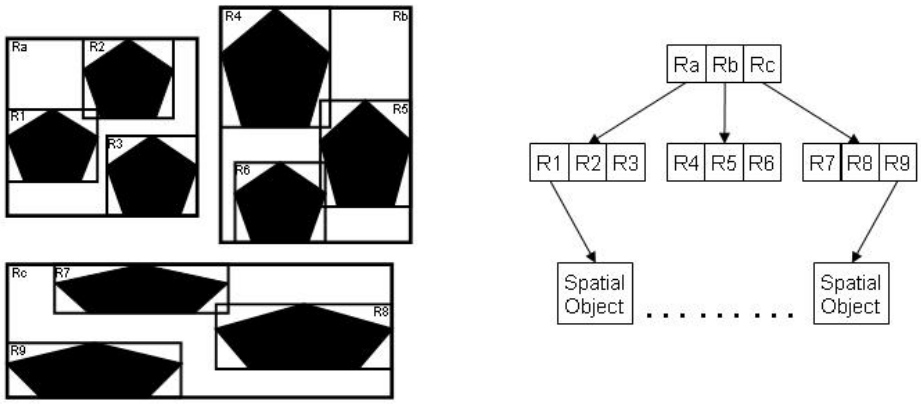


Figure 2. An R-tree example (adapted from [12]).

nodes (see R6 in Figure 2), a spatial search may include the visit of many nodes.

Starting from the original R-Tree structure, several improved versions have been proposed; when we move from spatial to spatio-temporal data, for instance, the temporal coordinate can be considered as an extra dimension and data can be indexed using three-dimensional R-trees [15]. Such an approach does not discriminate between spatial and temporal dimensions and is well suited for indexing the past, i.e., when only closed trajectories are considered.

STR-trees [4] extend R-tree with a different insert/split algorithm, while the characteristics of spatio-temporal data are captured by two access methods (STR-tree and TB-tree).

When objects' movements are constrained, for example on a network of connected road segments, a bidimensional R-tree can be used to index the static network's segments. In this case, each leaf contains a segment and a pointer to a monodimensional R-Tree that indexes the time intervals of objects' movements, as for FNR-Tree [16]. MON-tree [17] extends the FNR-tree by modeling the constrained network as a set of junctions and routes; a bidimensional R-tree is used to index polylines' bounding boxes while, for each polyline, another bidimensional R-tree indexes the time dimension of the objects within the polyline. PARINET [18] has been designed for historical data in constrained networks and models the network as a graph; trajectories are partitioned according to the graph partitioning theory. This method has been extended to handle continuous indexing of moving objects [19].

When dealing with real applications for indexing and querying large repositories of trajectories, the size of MBRs can be reduced by segmenting each trajectory and then indexing each sub-trajectory by using R-Trees; such an approach is described, for example, in [20], where a dynamic programming algorithm is presented for the minimization of the I/O for an average size query. SETI [21] segments trajectories and groups sub-trajectories into a collection of *spatial partitions*; queries run over the partitions that are most relevant for the query itself. TrajStore [22] co-locates on a disk block (or in a collection of adjacent blocks) trajectory' segments by using an adaptive multi-level grid; thanks to this method, it is possible to answer a query by only reading a few blocks.

All the above approaches, even presenting efficient solutions from different perspectives, are typically not supported in the available commercial products that make use of very efficient spatial indexes that, unfortunately, are typically restricted to the bi-dimensional case. For instance, PostGIS [23], a well known extension of PostgreSQL DBMS [24] for storing spatial data, even supporting three (and four)-dimensional data, does not support three-dimensional intersection and indexing operations. As a consequence, there is a strong interest in those methods which, even using off-the-shelf solutions, allow to solve the problem in the multi-dimensional space.

For this reason, in this work we propose a method able to index large sets of trajectories extracted from video cameras by means of off-the-shelf solutions; in our approach, we consider that objects are freely moving in our scene, without any kind of constraint. The main novelty lies in the fact that, differently from previous solutions, we do not suggest any new indexing three-dimensional structures. As a matter of fact, we propose a way to efficiently use available bidimensional solutions in order to solve the problem that spatial indexes for three-dimensional data are not widely available.

III. The Proposed Method

According to the line segment model, a trajectory T^k can be represented as a sequence of spatio-temporal points:

$$T^k =< P_1^k, P_2^k, ..., P_n^k >$$

with:

$$P_i^k = (x_i^k, y_i^k, t_i^k) \quad \forall i \in [1, n].$$

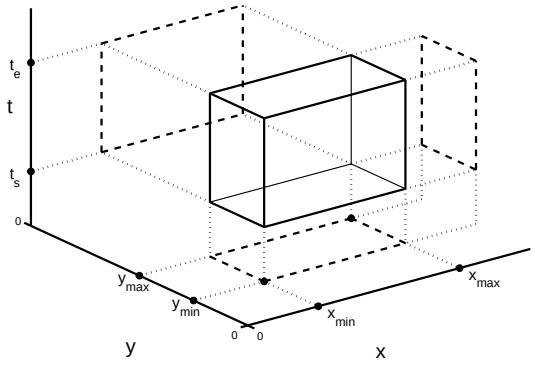


Figure 3. A query box representing a *TIQ*.

Each pair (x_i^k, y_i^k) refers to the spatial location of an object at the time instant t_i^k. As already mentioned, the trajectory is approximated by a polyline, each segment being the linear interpolant between two consecutive points (Figure 1).

A *Time Interval Query* (*TIQ*) aims at detecting all those trajectories passing through a given spatial area A in a given time interval $[t_s, t_e]$; assuming that the area A is rectangular, the latter is fully identified by two points in the xy plane $P_m^{xy} = (x_{min}, y_{min})$ and $P_M^{xy} = (x_{max}, y_{max})$. Each *TIQ* can be thus associated to a query box B:

$$B = \{(x_{min}, y_{min}, t_s), (x_{max}, y_{max}, t_e)\},$$

Differently speaking, the temporal dimension extends the rectangular area A in the 3D space (see Figure 3).

From a geometrical point of view, in order to solve a *TIQ* we need to find all those trajectories intersecting the query box B. A simple algorithm for retrieving the trajectories satisfying such a query is based on processing, for each trajectory, all its segments, starting from the first one: as soon as the intersection occurs, it can be concluded that the trajectory intersects the query box. In order to determine if a trajectory segment lies inside or outside a query box, a clipping algorithm can be used.

We propose to use the 2D Cohen-Sutherland Line Clipping Algorithm [25] (briefly summarized in Figure 4). According to it, the geometric plane is subdivided into nine areas by extending the edges of the query rectangle: if at least one of the segment endpoints lies inside the query box, the intersection is trivially verified (see segment AB in Figure 4.a); if, on the contrary, both the endpoints lie outside the query box, we check the position of the endpoints with respect to the query area: in some cases the intersection can be still trivially verified, as for CD and EF respectively in Figure 4.b, otherwise the segment is split at its intersection points and each obtained sub-segment is in turn inspected (as in the case of the segment GH and IL in Figure 4.c).

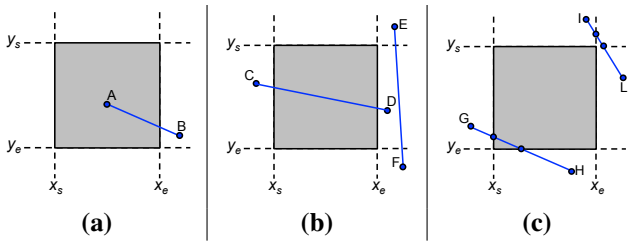


Figure 4. The Cohen-Sutherland Algorithm [25].

This clipping algorithm can be easily extended for dealing with 3D trajectories by considering 27 spatial regions, rather than 9.

Unfortunately, despite its simplicity, the use of a clipping algorithm is not suited for handling with large datasets, so demanding for more efficient approaches. In fact, the clipping algorithm has to process, for each trajectory, all its segments, starting from the first one until the intersection occurs. The worst case arises when a trajectory does not intersect at all the query box; in this case, in fact, all its segments must be processed, making this approach infeasible for large amount of data.

When the number of trajectories increases, more efficient approaches are thus mandatory; for example, suitable indexing strategies would be necessary to reduce the number of trajectories to be clipped. Although many spatial databases today available, both open source and commercial ones, provide very efficient spatial indexing techniques, these indexes schemes are unfortunately typically restricted to deal with 2D data; for this reason, it is worth trying to represent the actual 3D problem in terms of (one or more) 2D sub-problems, so as to fully exploit the efficient available bidimensional indexes.

A. The Solution Proposed in [1]

In [1], given a trajectory T^k and a query box B, we proposed a method according to which both T^k and B were first projected on the three coordinate planes; let T_{kz}^k and B_{kz} be the projections of T^k and of B on the kz plane. We then observed that, if a trajectory intersects the 3D query box, then each trajectory projection must also intersect the correspondent query box projection:

$$T^k \cap B \neq \emptyset \Rightarrow \left\{ \begin{array}{c} T_{xy}^k \cap B_{xy} \neq \emptyset \\ T_{xt}^k \cap B_{xt} \neq \emptyset \\ T_{yt}^k \cap B_{yt} \neq \emptyset \end{array} \right\} \quad (1)$$

Equation 1 represents a necessary but not sufficient condition, as the opposite is clearly not true. In fact, if all trajectory's projections intersect the correspondent box projection on the considered spaces, they do not necessarily intersect the 3D query box too. To better explain this concept, Figure 5.a shows, in the 3D space, a trajectory that does not intersect a given query box: it can be noticed that all the

trajectory projections intersect the correspondent query box projections (Figure 5.b-d).

Thus, as a matter of fact, if all projections of T^k intersect the correspondent box projections, we consider T^k as a candidate to be clipped in the three-dimensional space; the guess is that there will be not too many *false positives*.

According to the above considerations, in [1], for each three-dimensional trajectory T^k, we proposed to store three bi-dimensional trajectories obtained by projecting T^k on the xy plane (T^k_{xy}), on the xt plane (T^k_{xt}) and on the yt plane (T^k_{yt}).

Given a box B representing the time interval query to be solved, we similarly considered B_{xy}, B_{xt} and B_{yt}.

With this strategy, by using one of the available bi-dimensional indexes, it is possible to find on each plane the following three trajectory sets (Θ_1, Θ_2, Θ_3) in a very simple and efficient manner:

$$\Theta_{xy} = \{T_{xy} : MBR(T_{xy}) \cap B_{xy} \neq \emptyset\} \quad (2)$$

$$\Theta_{xt} = \{T_{xt} : MBR(T_{xt}) \cap B_{xt} \neq \emptyset\} \quad (3)$$

$$\Theta_{yt} = \{T_{yt} : MBR(T_{yt}) \cap B_{yt} \neq \emptyset\} \quad (4)$$

The set T of the trajectories candidate to be clipped in the 3D space is thus trivially defined as:

$$\Theta = \{T : T_{xy} \in \Theta_{xy} \wedge T_{xt} \in \Theta_{xt} \wedge \Theta_{yt} \in T_{yt}\} \quad (5)$$

This strategy, while taking advantage of widely available efficient bidimensional indexes, still presents two weak points. First, for a n points trajectory, we need to redundantly store $6{\cdot}n$ values ($2{\cdot}n$ for each of the three coordinate planes). Another subtle crucial point is that the use of bidimensional indexes is not optimized: as a matter of fact, the MBR of each projected trajectory can easily span a great percentage of the whole area.

In the following two subsections, the above problems will be separately handled and solutions for them will be presented.

B. Improving the Method by Removing Redundancies

It is possible to observe that, for a given trajectory T^k, rather than storing the three different trajectory projections in each coordinate plane, we can store T^k as the original sequence of points in the 3D space, and separately maintain three different bidimensional MBRs: $MBR_{xy}(T^k)$, $MBR_{xt}(T^k)$ and $MBR_{yt}(T^k)$. $MBR_{xy}(T^k)$ (respectively $MBR_{xt}(T^k)$ and $MBR_{yt}(T^k)$) is obtained by projecting on the xy (respectively xt and yt) plane the three-dimensional MBR of T^k.

It is worth noting that the redundancy introduced by the three MBR projections is not dependent on the number of points in the trajectory and, therefore, has only a marginal impact on the spatial complexity, since it only requires the storage of six pairs of points.

Assuming such a scheme, on each 2D plane we find the trajectories intersecting the corresponding 2D query box in a very efficient manner by using one of the available 2D indexes. Let Γ_{xy}, Γ_{xt} and Γ_{yt} be the resulting sets of trajectories defined as:

$$\Gamma_{xy} = \{T : MBR_{xy}(T) \cap B_{xy} \neq \emptyset\} \quad (6)$$

$$\Gamma_{xt} = \{T : MBR_{xt}(T) \cap B_{xt} \neq \emptyset\} \quad (7)$$

$$\Gamma_{yt} = \{T : MBR_{yt}(T) \cap B_{yt} \neq \emptyset\}, \quad (8)$$

where, as usual, B_{xy}, B_{xt} and B_{yt} are the projections of the 3D query box B. The set Θ of the trajectories candidate to be clipped in the 3D space is therefore now defined as:

$$\Theta = \Gamma_{xy} \cap \Gamma_{xt} \cap \Gamma_{yt} \quad (9)$$

Figure 6 resumes the method: given a set of trajectories and a query box (Figure 6.a), we discard the green trajectory since its $MBR_{xy}(T)$ and $MBR_{yt}(T)$ do not intersect the corresponding projection of the black query box (Figure 6.b and Figure 6.c). The candidate set Θ is thus composed by the other two trajectories, the red and the blue one, which are finally clipped, obtaining the desired output represented by the blue trajectory (Figure 6.e).

C. Optimizing the Selectivity of the 2D Indexes

It should be clear at this point that the entire system performance will strongly depend on the indexing phase and, as a consequence, on the capability to reduce the number of trajectories to be clipped in the three-dimensional space. At a more detailed analysis, the *selectivity* of the indexes in each plane is related to the area of the corresponding MBR which, in turn, only depends on the trajectory geometry, so being (apparently) fixed. This is the reason why we decided to introduce a segmentation stage, aimed at increasing the selectivity of the indexes.

Segmentation algorithms aim at subdividing each trajectory into consecutive smaller units, which we will refer to as trajectory units. We are interested in a segmentation algorithm able to exploit the characteristics of the available bi-dimensional indexes; this can be accomplished by decreasing the area of the projected MBR of each trajectory unit.

The proposed algorithm works recursively: initially (that is at iteration 0) it assumes that the trajectory T^k is composed by a single unit ${}^0U^k_1$, that is split into a set of m consecutive smaller units $\{{}^1U^k_1, \ldots, {}^1U^k_m\}$; each of the ${}^1U^k_i$ is in turn inspected and, if the stop criteria are not satisfied, it is further split.

Let us analyze how a generic unit ${}^{(i-1)}U = \{P_1, \ldots, P_m\}$ is split into $\{{}^iU_1, \ldots, {}^iU_n\}$; we first choose a *split-dimension* and a *split-value*. Assume, as an example and without loss of generality, that x has been chosen as the *split-dimension* and let x^* be the *split-value*. In addition, assume that $x_1 < x^*$. According to these hypotheses, ${}^iU_1$

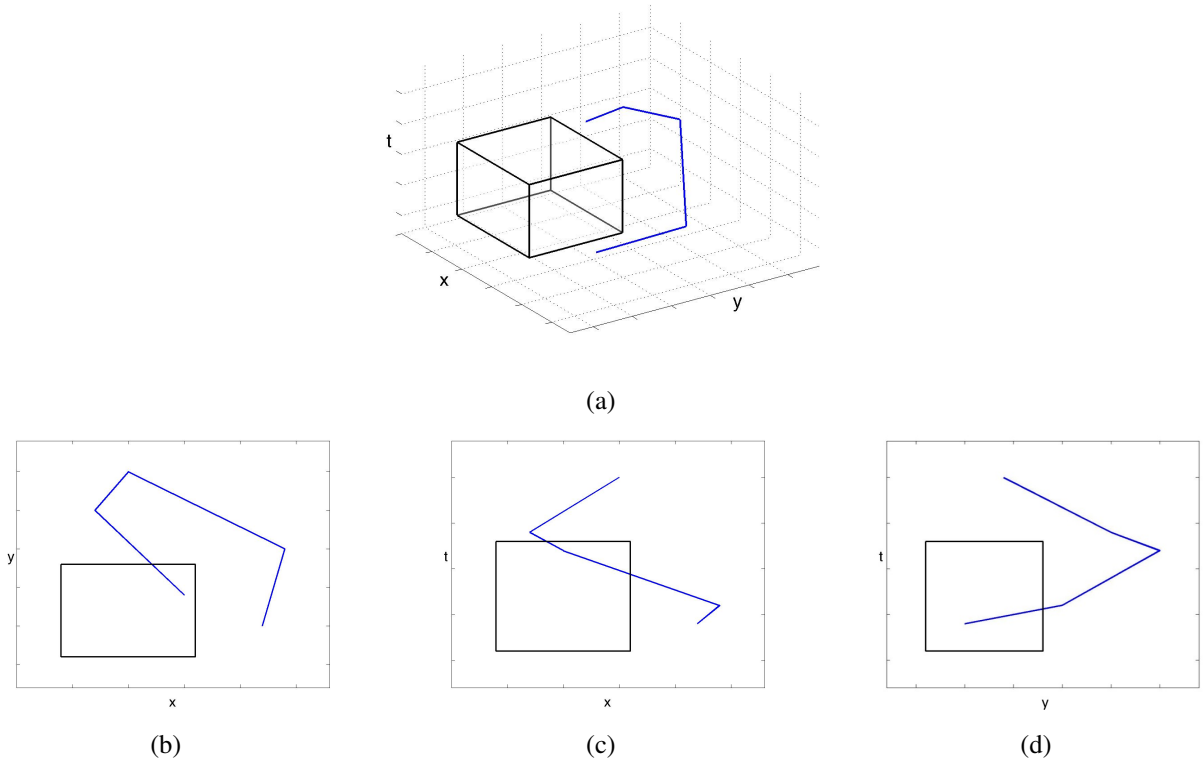


Figure 5. An example of 3D trajectory (a) and its projections on the different coordinate planes xy (b), xt (c) and yt (d). Although the trajectory does not intersect the query box, its projections do it.

is the set of the consecutive points lying on the left of the *split-value*:

$$^{i}U_1 = \{P_1, \ldots, P_k\} \tag{10}$$

where P_k is the first point such that $x_k \geq x^*$. Then, the second unit will be formed by the sequence of consecutive points lying on the right of the *split-value*:

$$^{i}U_2 = \{P_{k+1}, \ldots, P_l\} \tag{11}$$

where P_l is the first point such that $x_k \leq x^*$. The inspection of $^{(i-1)}U$ ends when the last point P_m is reached.

According to the above considerations, the criteria for the choice of the two parameters, *split-dimension* and *split-value*, play a crucial role. Since we aim at optimizing the indexing strategy, the proposed segmentation algorithm is based on the occupancy percentage on each 2D coordinate plane.

First we calculate the coordinate plane corresponding to the maximum among the three occupancy percentage values O_{xy}, O_{xt} and O_{yt} of the trajectory unit MBRs, with respect to the correspondent global volume of interest V:

$$O_{xy} = \frac{MBR^{xy}(U)}{V^{xy}} \tag{12}$$

$$O_{xt} = \frac{MBR^{xt}(U)}{V^{xt}} \tag{13}$$

$$O_{yt} = \frac{MBR^{yt}(U)}{V^{yt}} \tag{14}$$

Without loss of generality, suppose that the maximum occupancy percentage value is O_{xy} and, consequently, the corresponding plane is xy; let $width$ and $height$ be the two dimensions of $MBR_{xy}(U)$, respectively along the coordinates x and y; the *split-dimension* sd is defined as:

$$sd = \begin{cases} x & \text{if width} > \text{height} \\ y & \text{otherwise} \end{cases}$$

Given the *split-dimension* sd we choose, as the *split-value* sd^*, the MBR average point on the coordinate sd.

Figure 7 sketches the execution of the first iteration of our algorithm on the trajectory T (assumed to be composed at this iteration by a single unit U).

In Figure 7.a we are assuming that our volume of interest is:

$$0 \leq x \leq 150; 0 \leq y \leq 50; 0 \leq t \leq 30 \tag{15}$$

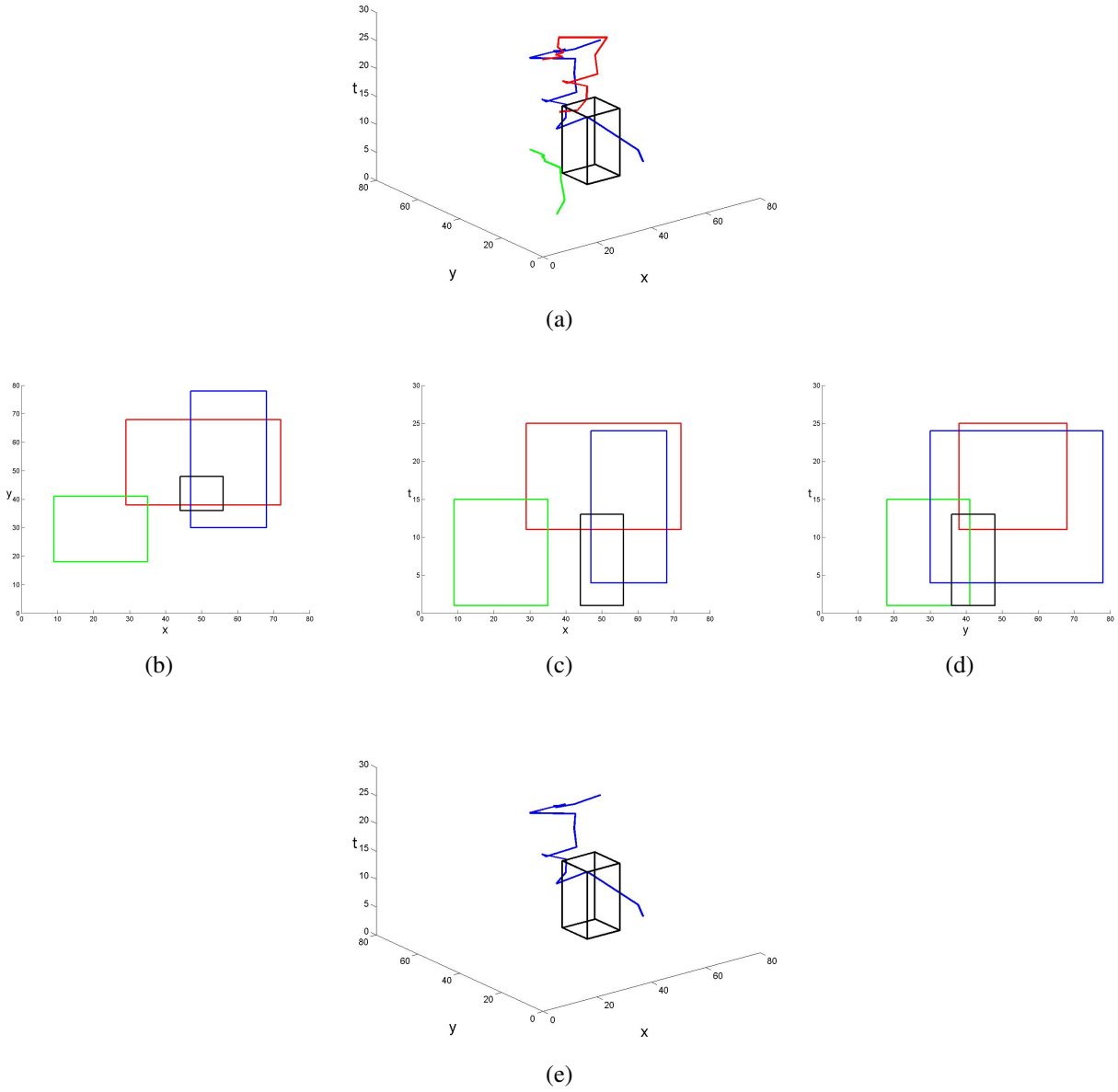


Figure 6. An overview of the method. (a) a query box and three trajectories; (b), (c), (d): the projections of trajectories' MBRs on the coordinate planes; (e) the final result of our method, after the application of the clipping algorithm on the blu and red trajectories.

We first consider $MBR_{xy}(U)$, $MBR_{xt}(U)$ and $MBR_{yt}(U)$ (Figure 7.b) obtaining that:

$$O_{xy} > O_{xt} > O_{yt}. \tag{16}$$

According to the above inequalities, we choose to operate on the xy plane. As the values of the dimensions are:

$$\text{width} = x_{max} - x_{min} = 135 \tag{17}$$

$$\text{height} = y_{max} - y_{min} = 40, \tag{18}$$

assuming that x has been chosen as the split-dimension, the split-value will be easily obtained as follows:

$$x^* = \frac{x_{max} + x_{min}}{2} = 72.5 \tag{19}$$

$$\tag{20}$$

Now we are in the position of segmenting the unit; Figure 7.c shows the current iteration, that is the segmentation of the unit while, in Figure 7.d, are shown, with different colors, the obtained 4 units with the corresponding MBRs. Last,

Figure 7.e shows the projections of the obtained MBRs on each coordinate plane.

The algorithm ends when all the trajectory units cannot be further subdivided, since at least one of the stop conditions has been reached for each unit; in particular, we employ two stop criteria . First, we choose not to segment trajectory units whose MBR areas are smaller than a fixed percentage of the entire scenario (PA^{min}); furthermore, we do not segment a unit with less than PS^{min} points.

Finally, a further refinement is needed in our algorithm for handling with the formation of the last unit; in fact, when the generic unit U^i is splitted, it can happen that the last unit is only composed by a few points. In this case, a suited strategy is introduced to specifically handle with this issue. Figure 8.a clarifies this concept: the last unit (the black one) is composed by three points, but LU^{min} is set to four. For this reason, this unit will be merged with the previous one (the green one, see Figure 8.b).

D. Summarizing the method

In this section we will summarize the proposed method. Each trajectory T^k is segmented (when it is loaded) and stored as a sequence of trajectory units $\{U_1^k, \ldots, U_l^k\}$; using one of the available bidimensional indexes, we select:

$$\Gamma_{xy}^U = \{U : MBR_{xy}(U) \cap B_{xy} \neq \emptyset\} \quad (21)$$

$$\Gamma_{xt}^U = \{U : MBR_{xt}(U) \cap B_{xt} \neq \emptyset\} \quad (22)$$

$$\Gamma_{yt}^U = \{U : MBR_{xy}(U) \cap B_{yt} \neq \emptyset\}, \quad (23)$$

where, as usual, B_{xy}, B_{xt} and B_{yt} are the projections of the 3D query box B.

The set Γ^U of units candidate to be clipped in the 3D space is therefore defined as:

$$\Gamma^U = \Gamma_{xy}^U \cap \Gamma_{xt}^U \cap \Gamma_{yt}^U. \quad (24)$$

By analyzing Γ^U we build, as follows, the set Θ:

$\Theta = \{\emptyset\}$
$\forall\ U^k \in \Gamma^U\ \{$
  if $k \notin \Theta\ \{$
    if U^k intersects B in the three-dimensional space $\{$
      $\Gamma = \Gamma \bigcup k$
    $\}$
  $\}$
$\}$.

The set Θ represents the result of a TIQ, if we are interested in the coordinate-based form, while the entire trajectories have to be extracted if we are interested in trajectory-based TIQs.

IV. Experimental Results

In order to characterize the efficiency of the proposed method, several trajectory-based *TIQs* were performed. We conducted our experiments on a PC equipped with an Intel quad core CPU running at 2.66 GHz, using the 32 bit version of PostgreSQL 9.1 server and the 1.5.3 version of PostGIS. Data have been indexed using the standard bidimensional R-tree over GiST (Generalized Search Trees) indexes; as the specialized literature confirms, this choice guarantees higher performance in case of spatial queries with respect to the PostGIS implementation of R-trees.

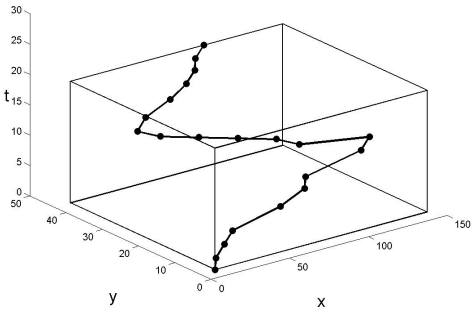
We represent each trajectory unit as a tuple:

$$(\underline{ID}, \underline{UID}, U, MBR_{xy}, MBR_{xt}, MBR_{yt}) \quad (25)$$

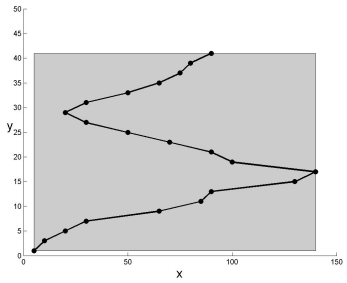
where ID is the moving objects identifier, UID identifies the trajectory unit, and U is the 3D trajectory unit, represented as a sequence of segments (a PostGIS 3D multi-line). Finally MBR_{xy}, MBR_{xt} and MBR_{yt} are the three unit's MBRs in each coordinate plane, xy, xt and yt respectively, represented as PostGis *BOX* geometries. Once data have been indexed, PostGIS provides a very efficient function to perform intersections between boxes and MBRs in a 2D space.

The experimental results have been obtained by testing the system performance on synthetic data, which have been generated as follows. Let W and H be the width and the height of our scene and S the temporal interval. Each trajectory T^i starting point is randomly chosen in our scene at a random time instant t_1^i; the trajectory length L^i is assumed to follow a Gaussian distribution, while the initial directions along the x axis and the y axis, respectively d_x^i and d_y^i, are randomly chosen. At each time step t, we first generate the new direction, assuming that both d_x^i and d_y^i can vary with probability PI_x and PI_y respectively; subsequently, we choose the velocity along x and y at random. The velocity is expressed in pixels/seconds and is assumed to be greater than 0 and less than two fixed maximum, V_x^{max} and V_y^{max}. Therefore, the new position of the object can be easily derived; if it does not belong to our scene, new values for d_x and/or d_y are generated. We refer to the scene populated with trajectories as the *Scenario*. Table I reports the free parameters and the values for the creation of the 30 different scenarios used in our experiments as well as the parameters used by the segmentation algorithm. Note that the worst case, corresponding to the maximum values of T and L, results in 10^4 trajectories with 10^4 points, for a total of 10^8 points to store and process; these values are over and above if compared with many real world datasets.

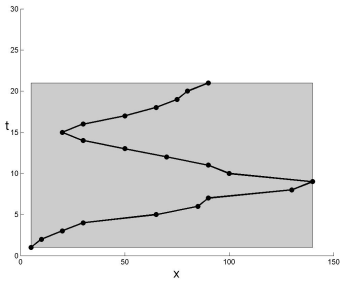
For the evaluation of the efficiency of the proposed segmentation algorithm, we generated and segmented 6000 trajectories with $L \in \{1000, 2000, 3000, 4000, 5000, 10000\}$; for each trajectory T^i we measured the number of obtained segments (N_{seg}^i) and the time needed to segment the trajectory (T_{seg}^i). Last, the obtained N_{seg}^i and T_{seg}^i are averaged over L, so obtaining $\overline{N_{seg}}(L)$ and $\overline{T_{seg}}(L)$. Figure 9 shows on the left the averaged number of segments ($\overline{N_{seg}}$) as L changes; not surprisingly we have, with very



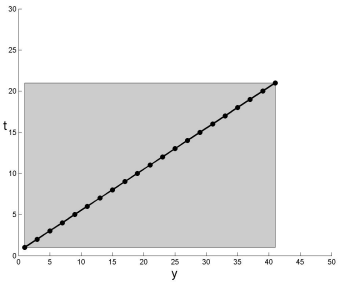
(a)



$O_{xy} = 72\%$

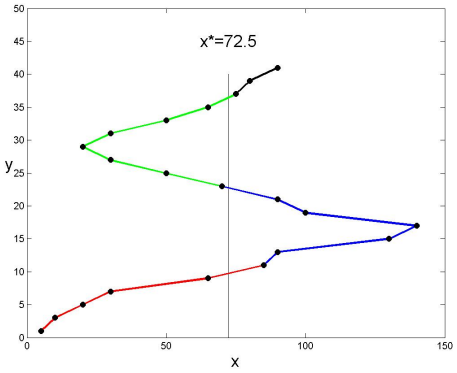


$O_{xt} = 69\%$

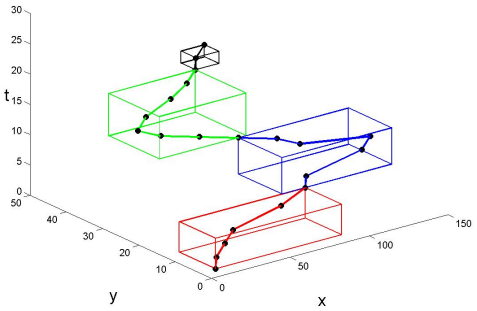


$O_{yt} = 53.4\%$

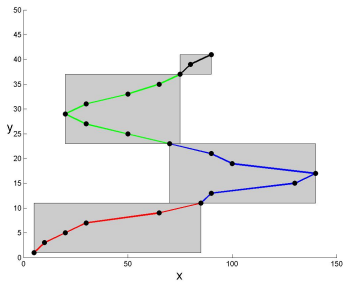
(b)

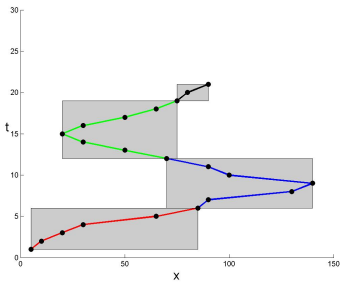


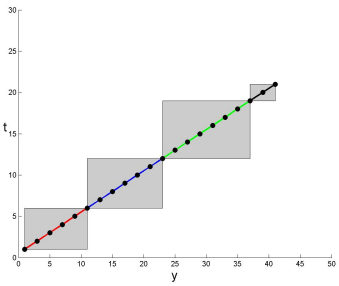
(c)



(d)

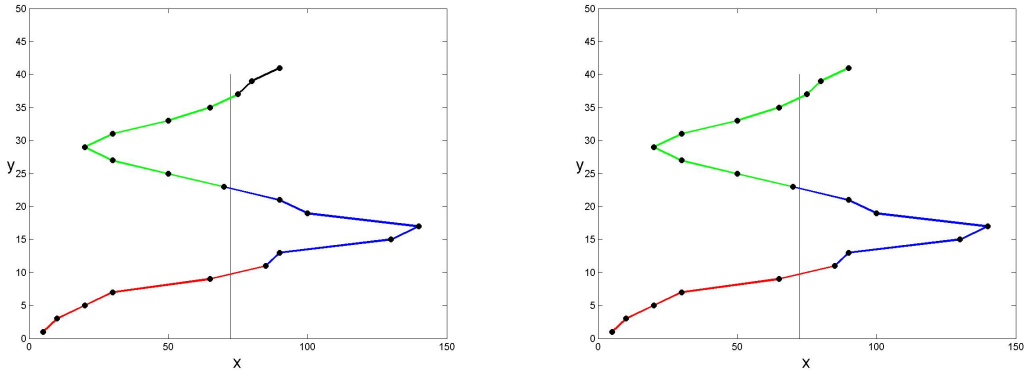






(e)

Figure 7. An overview of the segmentation algorithm. See text for details.



(a) (b)

Figure 8. The effect of LU_{min} on a segmented unit.

Table I
THE PARAMETERS USED IN OUR EXPERIMENTS.

Scene width (pixels)	10^4
Scene height (pixels)	10^4
Time interval length (secs)	10^5
Number of trajectories (T)	$\{1, 2, 3, 5, 10\} * 10^3$
Points in each trajectory (L)	$\{1, 2, 3, 5, 10\} * 10^3$
PI_x	5%
PI_y	5%
V_x^{max}	10 pixels/secs
V_y^{max}	10 pixels/secs
PA^{min}	1%
PS^{min}	100
LU^{min}	10

good approximation, that the number of segments linearly increases with L. On the right of Figure 9 is shown, in milliseconds, the averaged time $\overline{T_{seg}}$ needed to segment a trajectory with L points; with good approximation, $\overline{T_{seg}}$ quadratically increases with L.

The time needed to process a generic *TIQ* query (QT) is a function of at least 4 parameters, namely the number of trajectories T, the average trajectories' length L, the query cube dimension D_c, expressed as percentage of the entire scenario, and the position of the query box P_c:

$$QT = f(T, L, D_c, P_c). \quad (26)$$

Among the above parameters, P_c strongly influences the time needed to extract the trajectories as these are not uniformly distributed, especially in real world scenarios. In order to avoid the dependency on the query cube position, we decided to repeat the query a number of times inversely proportional to the query cube dimension, positioning the query cube in different positions, as shown in row N of Table II; finally, results are averaged to obtain:

$$\overline{QT} = f(T, L, D_c). \quad (27)$$

For the description of the experimental results we define three different set of experiments, obtained by fixing two

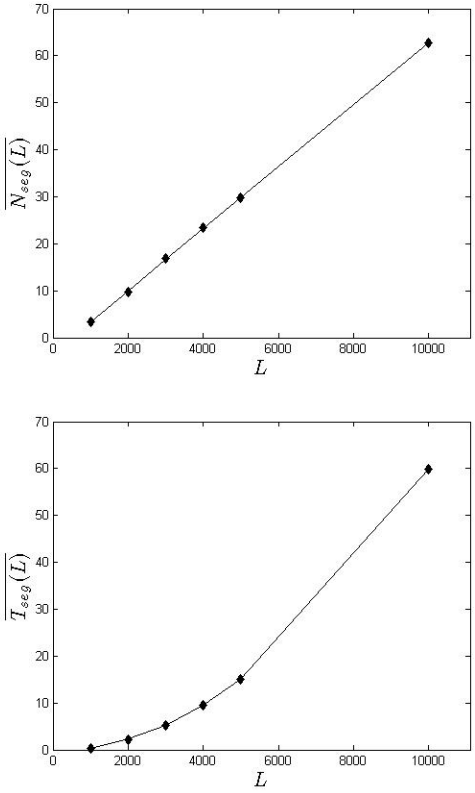


Figure 9. The performance of the segmentation algorithm.

Table II
NUMBER N OF TIMES EACH QUERY IS REPEATED AS D_c VARIES.

D_c	1%	5%	10%	20%	30%	50%
N	200	40	20	10	7	4

of the three parameters T, L and D_c and showing the variation of $\overline{QT}$ with respect to the third (free) parameter; an example is shown in Figure 10, which expresses the values

of $\overline{QT}$ with D_c and L fixed and T variable. Each diamond refers to the real value in seconds of $\overline{QT}$ for a given value of T; Figure 10 shows the curve which best interpolates the diamonds: in this case the approximation is linear, so obtaining a line. It is worth pointing out that, for the sake of readability, the figures for the experimental results will be expressed in a semi-log scale as, even not permitting to display part of the curves interpolating small values, it provides a greater comprehension of the system behavior for large values of the parameters.

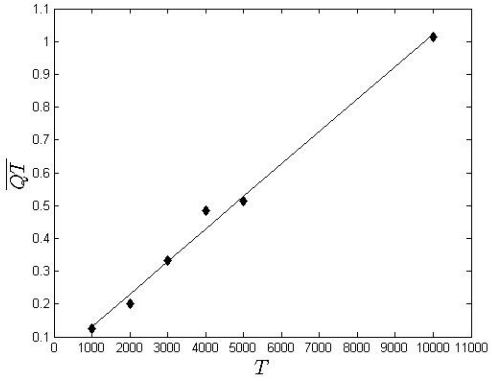


Figure 10. $\overline{QT}$ (in seconds) as the number of trajectories increases with $D_c = 5\%$ and $L = 5000$.

In Figure 11, $\overline{QT}$ relates to the variable number of trajectories T, for various values of D_c; the number of curves corresponds to the different fixed values of $L = \{1, 2, 3, 5, 10\} * 10^3$. The relationship between $\overline{QT}$ and T has been analyzed by polynomially approximating $QT(T)$: note that $\overline{QT}$ linearly increases with T, with a very small factor of approximation.

Diamond points in Figure 12 express $\overline{QT}$ in relation to the query box dimensions D_c and for $T = 3.000$ and $T = 10.000$; the number of curves corresponds to the different fixed values of $L = \{1, 2, 3, 5, 10\} * 10^3$. In this case we obtain that $\overline{QT}$ quadratically depends on D_c.

In Figure 13, the diamonds express $\overline{QT}$ as a function of L, for $D_c \in \{1\%, 5\%, 20\%, 30\%\}$, while the number of curves corresponds to the different fixed values of $T = \{1, 2, 3, 5, 10\} * 10^3$. Again we obtain that $\overline{QT}$ quadratically depends on L.

Finally, Figure 14 highlights the enhancement of the proposed indexing scheme as compared with the one presented in [1], for $D_c \in \{1\%, 20\%\}$. In particular, the diamonds refer to the new method, while the circles refers to the previous one. Note that, thanks to a novel indexing strategy and segmentation algorithm, the system performance significantly improved while the redundancy in stored data has been significantly removed.

V. Conclusion

In this paper we proposed an enhanced version of the retrieval system proposed in [1], aimed to index large amount of 3D trajectory data by using widely available 2D indexes. With respect to the previous method, two main improvements have been achieved: the former is the removal of the redundancy, obtained thanks to a novel indexing scheme; the latter is the optimization of the selectivity of the indexes, obtained by introducing a segmentation algorithm.

The experimental results, performed on synthetic data, show that the proposed solution is able to fully exploit the retrieving capabilities based on well established 2D indexes. In fact, the system performance have been significantly improved, if compared with the results presented in [1].

Further improvements in the performance will be achieved by applying the clipping algorithm in parallel to each candidate trajectory. This step can be easily implemented using multi-threading, in order to take advantage of multi-core and multi-processors systems. Strategies aiming at compressing data to be stored and retrieved are also being considered. Finally, we are extending our system in order to answer different query typologies.

Acknowledgment

This research has been partially supported by A.I.Tech s.r.l. (a spin-off company of the University of Salerno, www.aitech-solutions.eu)) and by the FLAGSHIP InterOmics project (PB.P05, funded and supported by the Italian MIUR and CNR organizations).

References

[1] A. d'Acierno, A. Saggese, and M. Vento, "A redundant bi-dimensional indexing scheme for three-dimensional trajectories," in *Proceedings of the First Conference on Advances in Information Mining and Management (IMMM 2011)*, 2011.

[2] A. d'Acierno, M. Leone, A. Saggese, and M. Vento, "A system for storing and retrieving huge amount of trajectory data, allowing spatio-temporal dynamic queries," in *Intelligent Transportation Systems (ITSC), 2012 15th International IEEE Conference on*, sept. 2012, pp. 989 –994.

[3] R. Di Lascio, P. Foggia, A. Saggese, and M. Vento, "Tracking interacting objects in complex situations by using contextual reasoning," in *Computer Vision Theory and Applications (VISAPP), 2012 International Conference on*, 2012.

[4] D. Pfoser, C. S. Jensen, and Y. Theodoridis, "Novel approaches in query processing for moving object trajectories," in *Proceedings of VLDB Conference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2000, pp. 395–406.

[5] D. Lim, D. Cho, and B. Hong, "Indexing moving objects for trajectory retrieval on location-based services," *IEICE - Trans. Inf. Syst.*, vol. E90-D, pp. 1388–1397, September 2007.

[6] M. F. Mokbel, T. M. Ghanem, and W. G. Aref, "Spatio-temporal access methods," *IEEE Data Eng. Bull.*, vol. 26, no. 2, pp. 40–49, 2003.

[7] Y. K. Huang and L. F. Lin, "Continuous within query in road networks," in *7th International Wireless Communications and Mobile Computing Conference (IWCMC)*, 2011, pp. 1176 – 1181.

[8] Y. Gao, B. Zheng, G. Chen, Q. Li, C. Chen, and G. Chen, "Efficient mutual nearest neighbor query processing for moving object trajectories," *Inf. Sci.*, vol. 180, pp. 2176–2195, June 2010. [Online]. Available: http://dx.doi.org/10.1016/j.ins.2010.02.010

[9] S. Shang, B. Yuan, K. Deng, K. Xie, K. Zheng, and X. Zhou, "PNN query processing on compressed trajectories," *GeoInformatica*, pp. 1–30, Oct. 2011. [Online]. Available: http://dx.doi.org/10.1007/s10707-011-0144-5

[10] Z. Li, Y. Gao, and Y. Lu, "Continuous obstructed range queries in spatio-temporal databases," in *System Science, Engineering Design and Manufacturing Informatization (ICSEM), 2011 International Conference on*, vol. 2, oct. 2011, pp. 267 –270.

[11] L.-V. Nguyen-Dinh, W. G. Aref, and M. F. Mokbel, "Spatio-temporal access methods: Part 2 (2003 - 2010)," *IEEE Data Eng. Bull.*, vol. 33, no. 2, pp. 46–55, 2010.

[12] Y. Manolopoulos, A. Nanopoulos, A. N. Papadopoulos, and Y. Theodoridis, *R-Trees: Theory and Applications (Advanced Information and Knowledge Processing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005.

[13] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *Proceedings ACM SIGMOD Conference*. New York, NY, USA: ACM, 1984, pp. 47–57.

[14] D. Papadias and Y. Theodoridis, "Spatial relations, minimum bounding rectangles, and spatial data structures," *International Journal of Geographical Information Science*, vol. 11, no. 2, pp. 111–138, 1997.

[15] Y. Theodoridis, M. Vazirgiannis, and T. Sellis, "Spatio-temporal indexing for large multimedia applications," in *Proceedings of the 3rd IEEE International Conference on Multimedia Computing and Systems*, 1996, p. 441 448.

[16] E. Frentzos, "Indexing objects moving on fixed networks," in *8th International Symposium on Advances in Spatial and Temporal Databases (SSTD 2003)*. Springer, 2003, pp. 289–305.

[17] V. T. De Almeida and R. H. Güting, "Indexing the trajectories of moving objects in networks*," *Geoinformatica*, vol. 9, pp. 33–60, March 2005. [Online]. Available: http://dl.acm.org/citation.cfm?id=1046957.1046970

[18] I. S. Popa, K. Zeitouni, V. Oria, D. Barth, and S. Vial, "Parinet: A tunable access method for in-network trajectories," in *Proceedings of the 26th International Conference on Data Engineering, ICDE 2010, March 1-6, 2010, Long Beach, California, USA*, F. Li, M. M. Moro, S. Ghandeharizadeh, J. R. Haritsa, G. Weikum, M. J. Carey, F. Casati, E. Y. Chang, I. Manolescu, S. Mehrotra, U. Dayal, and V. J. Tsotras, Eds. IEEE, 2010, pp. 177–188.

[19] I. Sandu Popa, K. Zeitouni, V. Oria, D. Barth, and S. Vial, "Indexing in-network trajectory flows," *The VLDB Journal*, vol. 20, pp. 643–669, Oct. 2011.

[20] S. Rasetic, J. Sander, J. Elding, and M. A. Nascimento, "A trajectory splitting model for efficient spatio-temporal indexing," in *Proceedings of the 31st international Conference on VLDB*. VLDB Endowment, 2005, pp. 934–945.

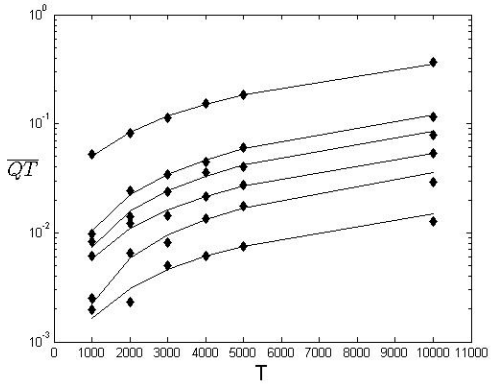
[21] V. P. Chakka, A. Everspaugh, and J. M. Patel, "Indexing large trajectory data sets with seti," in *CIDR*, 2003.

[22] P. Cudre-Mauroux, E. Wu, and S. Madden, "Trajstore: An adaptive storage system for very large trajectory data sets," *Data Engineering, International Conference on*, vol. 0, pp. 109–120, 2010.

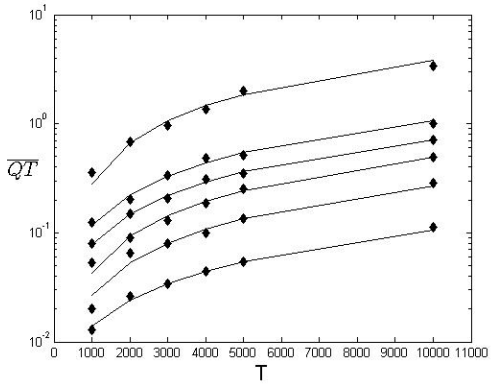
[23] R. Obe and L. Hsu, *PostGIS in Action*. Greenwich, CT, USA: Manning Publications Co., 2011.

[24] PostgreSQL Global Development Group, "PostgreSQL," Accessed: 12/19/2012. [Online]. Available: http://www.postgresql.org/

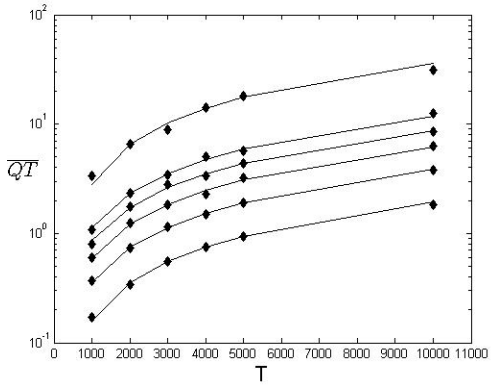
[25] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice in C (2nd Edition)*. Addison-Wesley, 2004.



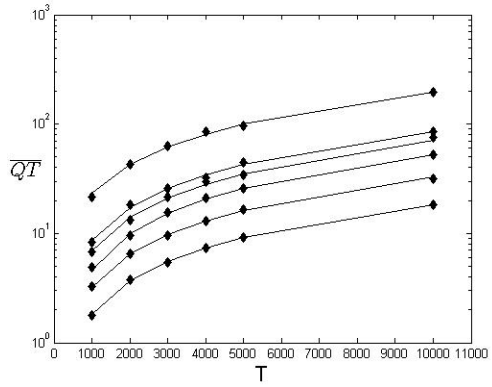
$D_c = 1\%$



$D_c = 5\%$

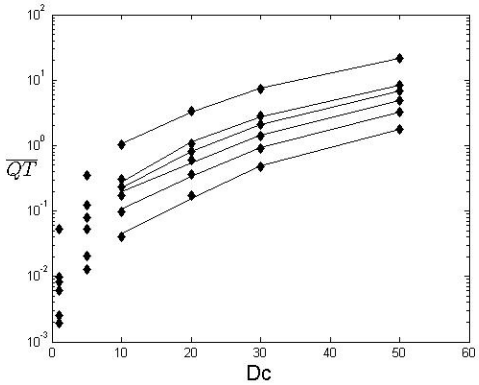


$D_c = 20\%$

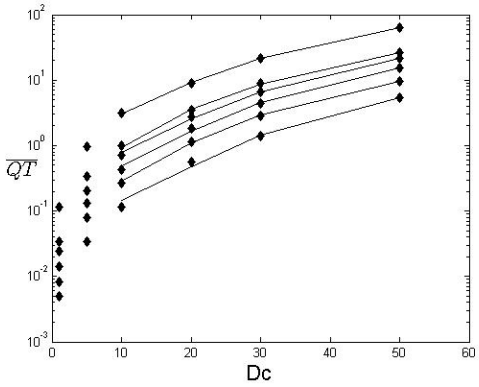


$D_c = 50\%$

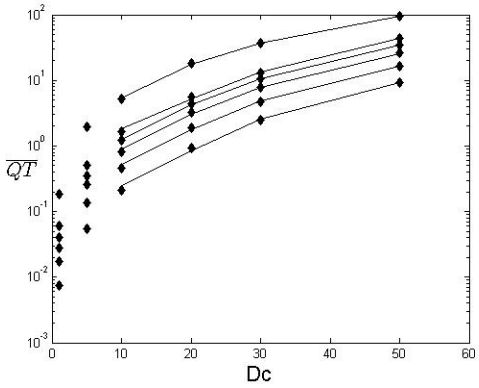
Figure 11. $\overline{QT}$ (in seconds) as the number of trajectories increases having the number of points in each trajectory as parameter.



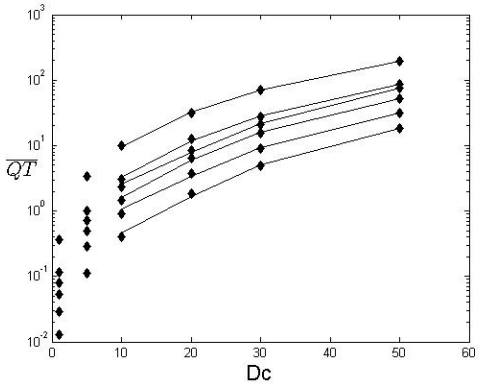
$T = 1000$



$T = 3000$

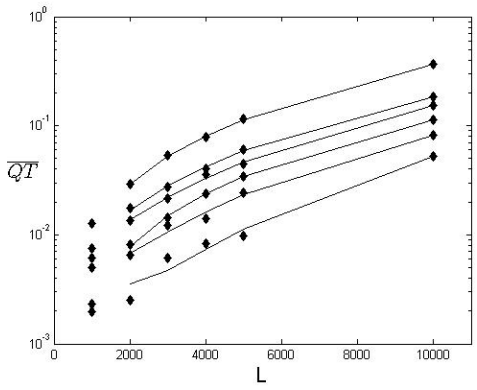


$T = 5000$

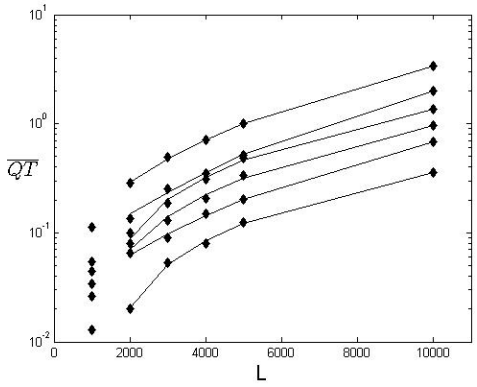


$T = 10000$

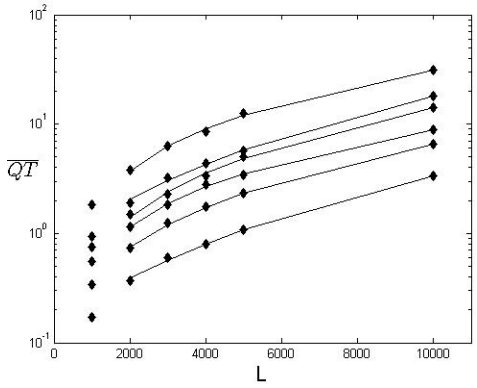
Figure 12. $\overline{QT}$ (in seconds) as the dimension of the querying cube (in percentage of the whole volume) increases and having L as parameter.



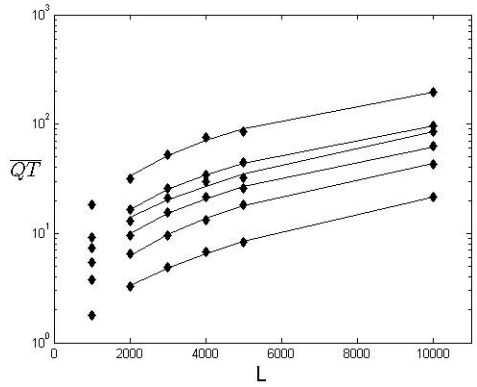
$D_c = 1\%$



$D_c = 5\%$

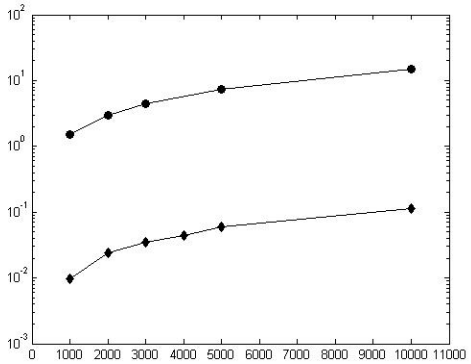


$D_c = 20\%$

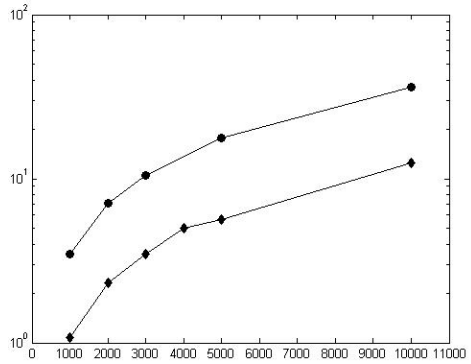


$D_c = 50\%$

Figure 13. $\overline{QT}$ (in seconds) as the number of points in each trajectory increases and having the number of trajectories as parameter.



$D_c = 1\%$



$D_c = 20\%$

Figure 14. The results obtained with the solution proposed in [1] (circles) compared with the results obtained with the solution here as T varies (L=5000).

Multiple Similarities for Diversity in Recommender Systems

Laurent Candillier*, Max Chevalier†, Damien Dudognon*†, and Josiane Mothe†‡

* *OverBlog, Ebuzzing, Toulouse, France*
Email: firstname.name@ebuzzing.com
† *IRIT, UMR5505, CNRS, Université de Toulouse, France*
Email: firstname.name@irit.fr
‡ *IUFM, Université de Toulouse, France*
Email: firstname.name@univ-tlse2.fr

***Abstract*—Compared to search engines, recommender systems provide another means to help users to access information. Recommender systems are designed to automatically provide useful items to users. A new challenge for recommender systems is to provide diversified recommendations. In this paper, we investigate an approach to obtain more diversified recommendations using an aggregation method based on various similarity measures. This work is evaluated using three experiments: the two first ones are lab experiments and show that aggregation of various similarity measures improves accuracy and diversity. The last experiment involved real users to evaluate the aggregation method we propose. We show that this method allows the balance between accuracy and diversity of recommendations.**

Keywords-Recommender System; Diversity; Similarity Measures; Users Study; Information Retrieval

I. INTRODUCTION

As explained by Ricci *et al.* [2], "Recommender Systems (RS) are software tools and techniques providing suggestions for items to be of use to a user". RS are uselly classified according to how item suggestions are generated; three categories are genearlly distinguished [3], [4]:

- Collaborative filtering that uses social knowledge to generate recommendations;
- Content-based filtering that uses content features to generate recommendations;
- And hybrid filtering that mixes content-based and collaborative filtering approaches.

Kumar and Thambidurai underline in [5] that "recommender systems are characterized by cross-fertilization of various research fields such as: Information Retrieval, Artificial Intelligence, Knowledge Representation, Discovery and Data/Text Mining, Computational Learning and Intelligent and Adaptive Agents". Indeed, when considering content-based filtering techniques, an important issue is to match various items and to identify those that should be recommend to a given user. In content-based R, such a matching is mainly based on similarity measures coming from Information Retrieval (IR) field [6].

IR usually sorts the retrieved documents according to their similarity with the user's query [7]. Doing so, IR systems assume that document relevance can be calculated independently from other documents [8]. As opposed to this assumption, various studies consider a user may prefer to get documents treating of various aspects of her information need rather than possibly redundant aspects within documents [9], [10]. Diversity pursues this goal.

Document diversity has many applications in IR. First, it is considered to be one solution to query term ambiguity [8]. Indeed, queries as expressed by users are not enough to disambiguate terms. To answer ambiguous queries, Clarke *et al.* [8] suggest that IR can provide the user with a range of documents that corresponds to the various term senses. In that case, redundancy can be penalized by lowering the rank of a document that is too similar to a document ranked higher in the list. Following the same idea and to face query term ambiguity, Chifu and Ionescu propose a non-supervised method that clusters documents and re-order retrieved documents based on the clustering results [11].

Diversity became a real challenge in RS field too [12]. It aims at tackling at least two objectives: removing redundancy in the recommendation list (i.e. avoiding recommendation of items that are too similar) and taking into account diverse interests.

In the literature two kinds of diversity have been proposed: individual diversity and aggregate diversity [13]. Individual diversity aims at recommending to a single user some recommendations that are not redundant; aggregate diversity aims at recommending items that are not redundant from one user to another (considering the "long tail" phenomenon). This paper focuses on individual diversity to provide a user with a diversified list of recommendations.

In order to achieve this goal, we investigate the relation between diversity and similarity measures. We study how different similarity measures, based on various aspects of recommended items, can be aggregated to provide more diversified recommendations while keeping a good accuracy.

Indeed, our main objective being to consider the variety of the users' expectations, the recommended items must be sufficiently diversified to cover a large range of expectations. This intuition comes from the fact that item relevance may

be multi-dimensional and dynamic [14]. This idea was initially developed by Candillier *et al.* [1] and is extended in this paper.

The paper is organized as follows: Section II presents the related works dealing with the links between similarity measures and diversity. We describe in Section III two first experiments based on TREC [15] IR tasks (*adhoc* and *diversity*). These experiments show that aggregation of various similarity measures may improve accuracy and diversity. In Section IV, we complete these experiments with a user study on a blog platform consisting of more than 20 million of articles. We show the positive impact of the aggregation of various similarity measures on the users' perception of diversity in recommendations. Section V concludes this paper and underlines our future work.

II. Related works

In this section, we explain that diversity can result from the use of various similarity measures (notice that similarity measures used in RS mostly come from IR).

Users' interests are different, multidimensional and dynamic [14]. This assumption is confirmed forasmuch as document usefulness can be estimated differently. Mothe and Sahut [16] consider that a document can be evaluated on various criteria:

- Relevance;
- Information validity;
- Physical and ergonomic aspects.

Each of these criterium being in turn depicted by several sub criteria.

To deal with the variety of interests, IR systems diversify the retrieved documents [17], [12]. Doing this, the systems maximize the chances of retrieving at least one relevant document to the user [18].

IR literature distinguishes topicality and topical diversity. Topicality makes reference to which extent the document may be related to a particular topic [19] and is not related to diversity. Topical diversity refers both to extrinsic diversity and intrinsic diversity. The former helps to dispel the uncertainty resulting from the ambiguity of the user's needs or from the lack of knowledge about user's needs [20]. The intrinsic diversity, or novelty, intends to eliminate redundancy in the retrieved documents [8]. Very similar documents allow the system to increase the accuracy but do not improve the user's satisfaction [21]. The intrinsic diversity allows the system to present to the user:

- Various points of view;
- An overview of the topic that can only be achieved by considering simultaneously several documents;
- Or even to check the information reliability [20].

Topical diversity is generally used to reorder the retrieved documents. Two types of methods are generally used. The first one considers the reordering process as a clustering problem, while the other is based on a selection method such as the Maximal Marginal Relevance (MMR) proposed in [9].

With regard to clustering method, He *et al.* [22] use Single Pass Clustering (SPC). In this approach, the first document in the result list is selected and assigned to the first cluster. Then, the algorithm processes down the list of retrieved documents and assigned each document to the nearest cluster. If the document-cluster similarity is below a defined threshold, the document is assigned to a new cluster. Bi *et al.* [23] obtained better results using the k-means algorithm [24]. Whatever the algorithm used, assignment to different clusters is generally done using a distance such as the Euclidean distance or the Cosine measure, eventually weighted by the terms frequency. Meij *et al.* [22] apply a hierarchical clustering algorithm on the top fifty retrieved documents using document language modeling based approach. The document selection phase used to build the result list is based on cluster quality and stability metrics. Then, the best documents from each cluster are selected.

In these approaches, the clustering step takes place after a set of documents has been retrieved; the documents are grouped together according to the sub-topics clusters identify.

Topical diversity is also used to reduce redundancy in the retrieved document list. MMR [9] or sliding window approaches [25] aim at selecting the documents maximizing the similarity with the query and, at the same time, minimizing the similarity with all the documents already selected. The function used to compute the similarity between a given document and the documents already selected can differ from the similarity function used to estimate the relevance with the query [9].

Several approaches select the documents to be reordered using indicators or filters to increase the diversity in the results. Kaptein *et al.* [26] employ two types of document filters: a filter, which considers the number of new terms brought by the document to the current results and a link filter, which uses the value added by new input or output links to select new documents. Furthermore, Ziegler *et al.* [21] propose an intra-list similarity metric to estimate the diversity of the recommended list. This metric uses a taxonomy-based classification.

However, some user's needs cannot be simply satisfied by topic-related documents. For instance, serendipity aims at bringing to the user attractive and surprising documents she might not have otherwise discovered [27]. It is an alternative to topical diversity. For example Lathia *et al.* [28] investigate the case of temporal diversity. In the other hand, Cabanac *et al.* [29] consider organizational similarity that considers how the users sort their documents in a folder hierarchy.

Thus, similarity measures are different and may either be based on document content or structure, or on document usage considering popularity or collaborative search.

In the literature, several types of similarity functions have been considered:

- Based on document content: to be similar two documents should share indexing terms. Example of such measures are the Cosine measure [7], or semantic measures [30], [31];
- Based on document popularity such as the BlogRank [26];
- Collaborative: the document score depends on the scores that previous users assigned to it [32];
- Based on browsing and classification: document similarity is either based on browsing path [33] or considering the categories users assigned to viewed documents [23];
- Based on relationships: social similarity functions use relationships between contents and users [34], [35].

In this context, we hypothetise that diversification of recommendations can be obtained by combining several similarity metrics. The reason is that each metric answers a specific need or represents a particular view of the information interest. Similarly, Ben Jabeur *et al.* [34] combined a content similarity measure, based on TF-IDF [36], with a social measure which reflects the relationships in a social network of authors. The main difficulty with this kind of approaches lies in the way of combining the similarity measures. Whether it is a linear combination, or a successive application of measures, a combination boils down to give some importance to each measure and to favor certain facets over others.

An alternative to similarity combination is to consider different similarity metrics independently. Amazon.com [37] offers several recommendation lists to the user and indicates the type of measure used in naming these lists (e.g. "Customers who viewed this item also bought", "Inspired by the general trends of your purchases"). However, this independence of similarity metrics sometimes leads to a redundancy of information: one document can be recommended to the user in several lists of recommendations.

Fusion approaches offer a way to solve this problem. Indeed, the fusion of results from different similarity metrics within a single list of recommendations eliminates duplicates. Shafer *et al.* [18] and Jahrer *et al.* [38] propose to merge multiple sources of recommendations and therefore present a "Meta RS". Depending on the fusion approach, it is possible to favor documents appearing in multiple lists or not [39].

Finally, a graph approach can be used to fuse a set of similarity measures [40]. The results of each measure help to establish links between documents. These links are materialized by edges in a graph, weighted by the similarity scores and the documents are represented by nodes. The number of edges between two documents is only limited by the number of similarity measures used.

To be able to evaluate and compare topical diversity oriented approaches, TREC Web 2009 campaign [15] defines a dedicated topical diversity task. This task is based on the ClueWeb09 dataset, which consists in roughly one billion web pages crawled during January and February 2009. The size of the whole dataset (named Category A) is about twenty five Terabytes of data in multiple languages. The set B of the corpus we use for our experiments only focuses on a subset of English-language documents, roughly fifty million documents. The *diversity* task uses the same fifty queries as the *adhoc* tasks [41].

Clarke *et al.* [42] present the panel of metrics used to estimate and compare the performances of the topical diversity approaches. In our experiments, we only consider the Normalized Discounted Cumulative Gain (α-nDCG) [8] which is the metric used for the TREC Web 2009 evaluation campaign.

This evaluation framework is not enough to evaluate RS diversity when not only content-based elements are used but others also. Indeed, it turns out that the proposed approaches, either based on clustering algorithms or on selection criteria, are mainly focused on content and on topical diversity. The available evaluation frameworks, such as the TREC Web *diversity* task, have been designed to measure the performances of these content-based approaches. To be able to evaluate other types of diversity, like serendipity, and to truly gauge the user's satisfaction, a user study is necessary [43].

The hypothesis of our work is that diversity obtained when aggregating the lists of items resulting from different similarity measures is a means to diversify recommendations in a RS. Indeed, even if a unique recommendation method is efficient in the majority of the cases, it is useful to consider other users' expectations. Content-based diversity, but also other sorts of diversity, should be considered in recommendations.

This paper aims at showing the impact on diversity in RS of an aggregation method applied to various similarity measures. To achieve this goal we propose to verify three hypotheses:

- The aggregation of similarity measures considering the same aspect of item (e.g. item topic) improves the accuracy of recommended items;
- The aggregation of a variety of similarity measures improves the overall diversity of recommended items;
- The users' perception of diversity is high when aggregating various similarity measures while keeping a perception of a good level of accuracy.

These hypotheses are studied in this paper through three experiments we conducted.

III. Experiments

We hypothesize there is not one single approach that can satisfy the various users' expectations, but a set of

complementary approaches. In our view, each approach could correspond to a different point of view on the information and thus answers to specific users' expectations. We hypothesize that aggregating various approaches could be a relevant solution. To start with, we decided to verify that two distinct approaches retrieve different documents for a given IR tasks (*adhoc*, *diversity*). We then show the positive impact of the aggregation of these distinct approaches on accuracy.

For the experiment, we consider several systems, which were evaluated within the same framework to ensure they are comparable, and for which the evaluation runs were available. We focus on the *adhoc* and *diversity* tasks of the TREC Web 2009 campaign considering only the set B of the corpus to get comparable systems. Moreover, we choose the four best runs for each task rather than taking into account all the submitted ones.

To compare the selected runs, we follow the framework and the metric proposed by Lee [44] in the context of IR. This framework is widely used in the literature. We compute the overlap for each pair of runs, that is to say the number of common documents between the two compared runs. The overlap is computed for the n first documents. We first compare the global overlap considering all retrieved documents. Then, we focus on the relevant document overlap and on the non relevant document overlap.

We use the metric proposed by Lee [44] and defined as follows:

$$overlap = \frac{2 \cdot |run_1 \cap run_2|}{|run_1| + |run_2|} \quad (1)$$

Where run_1 and run_2 are the documents of the two runs to be compared. The value of the overlap is between 0, when both runs have no common document, and 1 if the same documents are retrieved by run_1 and run_2.

In this section, we compare the results obtained by the best runs in two tasks: *adhoc* task and *diversity* task.

A. Adhoc task experiment

1) Adhoc task and compared runs: The TREC *adhoc* task is designed to evaluate the performances of systems, that is to say their ability to retrieve relevant documents for a given query. These systems have to return a ranked list of documents from the collection, ordered by decreasing expected relevance. The expected relevance considers each document independently: it does not take into account the other documents that appear before it in the retrieved list. The full evaluation counts fifty queries [41].

The performances of the different evaluated systems are compared using MAP which is based on precision. The precision P defines the proportion of relevant documents among the retrieved documents and is formally expressed by:

$$P = \frac{|\{relevant\ documents\} \cap \{retrieved\ documents\}|}{|\{retrieved\ documents\}|} \quad (2)$$

Thus, the average precision $AveP$ is defined as:

$$AveP = \frac{\sum_{k=1}^{n} (P(k) \cdot rel(k))}{|\{relevant\ documents\}|} \quad (3)$$

Where

- $P(k)$ is the precision considering the k first documents in the result list;
- $rel(k)$ is a function which indicates if a document is relevant (1) or not (0).

Finally, the MAP measure, used for the TREC evaluation campaigns, is the mean of the average precision scores $AveP$ for each query q of the set of queries Q:

$$MAP = \frac{\sum_{q=1}^{|Q|} AveP(q)}{|Q|} \quad (4)$$

The scores of the four best run at the TREC Web 2009 *adhoc* task are presented in Table I.

Table I
TREC WEB 2009 ADHOC TASK RESULTS

Group Id	Run Id	MAP
UDel	udelIndDRSP	0.2202
UMD	UMHOOsd	0.2142
uogTr	uogTrdphCEwP	0.2072
EceUdel	UDWAxQEWeb	0.1999

The runs (Run Id) we kept are the following ones:

- udelIndDRSP: this run, generated using the Indri search engine [45], combines the query-likelihood language model with the Markov Random Fields (MRF) model of term dependencies and the pseudo relevance feedback with relevance models. It also uses a metric to define trust in a domain. This metric is supported by content filtering whitelists and blacklists and publicly-available sendmail [46];
- UDWaxQEWeb: relies on an axiomatic retrieval framework where the relevance is modeled with retrieval constraints. The retrieval process aims at searching for functions that can satisfy these constraints. The approach is completed by a query expansion step. The expansion terms semantically related to the query terms are extracted from a Web search engine [47];
- UMHOOsd: uses a model based on the MRF in a distributed retrieval system built on Hadoop [48], the open source implementation of MapReduce [49];
- uogTrdphCEwP: uses the Terrier IR platform [50] with the implementation of the DPH weighting model derived from the Divergence From Randomness (DFR)

model. A query expansion step completes the retrieval process using the ClueWeb09 Wikipedia [51] documents [52].

2) Results:

a) Overlap of retrieved documents: Figure 1 presents the average overlap and precision for the four runs selected in the *adhoc* experiment, considering the fifty queries of the task. The precision and the overlap both take their values in between 0 and 1. We note that when we focus only on the first retrieved documents the global overlap is low, in spite of the fact that the first retrieved documents are most relevant. For example, taking the ten first documents for which the precision reaches its highest value (0.386), the average global overlap is only 0.255. The global overlap is low, even on a set of hundred documents (0.390).

Table II
GLOBAL OVERLAP CONSIDERING THE RUNS OF THE TREC WEB 2009 ADHOC TASK

Runs		udwa	umhoo	udel
umhoo	Relevant	0.8120		
	Non Relevant	0.5958		
udel	Relevant	0.7616	0.7806	
	Non Relevant	0.4721	0.5177	
uog	Relevant	0.7223	0.7583	0.6915
	Non Relevant	0.5133	0.4754	0.4066
Average	Relevant		0.7544	
	Non Relevant		0.4968	

Next, we focus on the average global overlap of relevant and non-relevant documents. We first compute the overlap (see Table II) for the overall runs, that is to say considering one thousand documents per query. We obtain an average global overlap equals to 0.754 for the relevant documents, and 0.497 for the non-relevant ones. These results are consistent with Lee's conclusions [44] on the TREC3 *adhoc* task: different runs retrieve different non-relevant documents but retrieve similar relevant documents.

Generally speaking, IR users focus on the first documents only [53]. In the same way, in the context of RS, only a small set of recommendations is proposed to the user. The choice is harder when there are a lot of documents provided to the user [54].

Therefore, we further analyze the evolution of relevant and non-relevant document overlap depending on the number of retrieved documents. Figure 2 shows that when we consider at the fifty top documents, the overlap is low for both relevant and non-relevant documents and it is pretty much the same until twenty documents.

b) Aggregating retrieved documents: The experiment demonstrates that, for a given query, two distinct systems are unlikely to retrieve the same set of top documents. Therefore, it is reasonable to expect that system result aggregation is relevant and could help to improve the accuracy of the

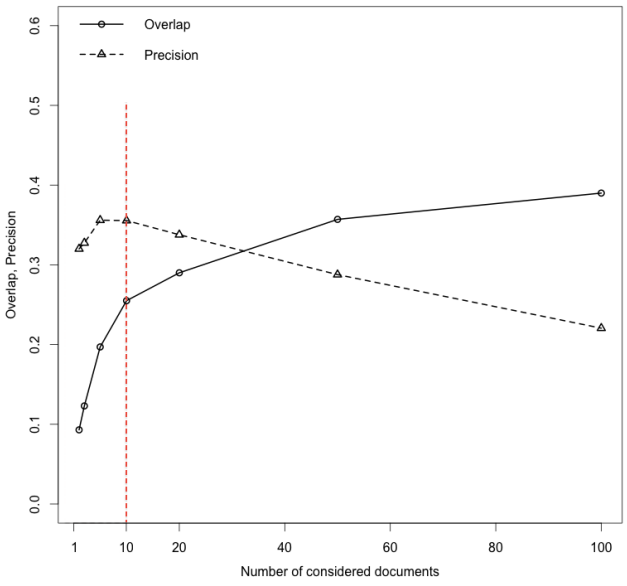


Figure 1. Average global overlap and precision for TREC Web 2009 adhoc task

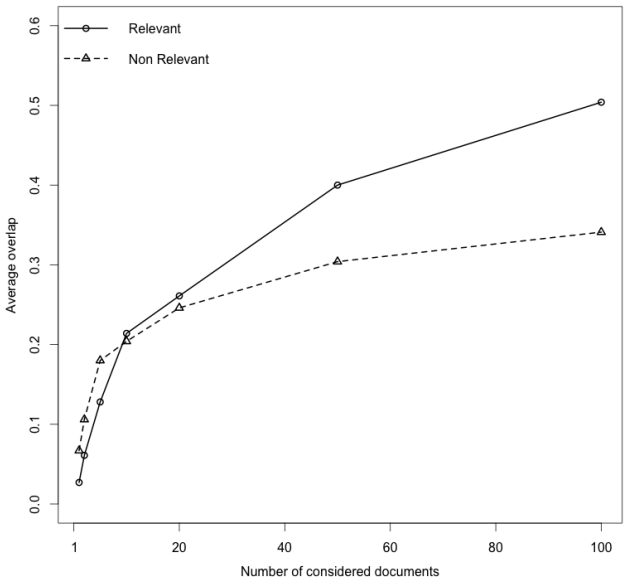


Figure 2. Average overlap for TREC Web 2009 adhoc task considering relevant and non relevant documents

results. To assess the relevance of approach aggregation, we aggregate the four runs previously used. For each query, all the retrieved document sets are aggregated using the fusion CombMNZ function [55] to generate a new run. CombMNZ has shown to be very efficient in the IR context.

$$CombMNZ(d_i) = (\sum_{j=1}^{n} w_{ij}) \cdot Count(d_i) \qquad (5)$$

Where

- d_i is a document;
- n is the number of similarity measures ;
- w_{ij} is the document score obtained with a similarity measure;
- $Count$ is a function which indicates the number of similarity measures that have retrieved the document d_i.

Then, following the *adhoc* task evaluation framework, we compute the MAP and obtain a score of 0.237, which outperforms the best run (0.2202).

In the next experiment, we apply the framework used on the *adhoc* task to the *diversity* task. We first aim at checking if various diversity-oriented approaches retrieve the same relevant documents. Subsequently, we investigate the consequences of approach aggregation on the diversity in the retrieved documents.

B. Diversity task experiment

1) Diversity task and compared runs: Similarly to the previous experiment, we center on several systems submitted at the TREC Web *diversity* task. All these systems aim at providing users with diversified result lists. The goal of the *diversity* task is to retrieve a ranked set of documents that together provide complete coverage for a query. Moreover, excessive redundancy should be avoided in the result list. The probability of relevance of a document depends on the documents that appear before it in the result list [41]. The queries are the same for the *adhoc* and the *diversity* tasks. The evaluation measures and the judging process differ from the *adhoc* task. The measure used for the TREC Web 2009 *diversity* task is the α-*nDCG* [8] derived from the Discounted Cumulative Gain (DCG) proposed in [25].

The DCG is based on the gain vector G and on the Cumulative Gain CG defined as:

$$G[k] = \sum_{i=1}^{m} J(d_k, i)(1-\alpha)^{r_{i,k-1}} \quad (6)$$

$$CG[k] = \sum_{j=1}^{k} G[j] \quad (7)$$

Where $J(d_k, i)$ is equal to 1 if the kth document is judged as relevant and 0 otherwise. Thus, DCG is formalized by:

$$DCG[k] = \sum_{j=1}^{k} \frac{G[j]}{log_2(1+j)} \quad (8)$$

Normalized Discounted Cumulative Gain ($nDCG$) is the ratio between the Discounted Cumulative Gain DCG and the ideal Discounted Cumulative Gain DCG':

$$nDCG[k] = \frac{DCG[k]}{DCG'[k]} \quad (9)$$

For the evaluation process, α is set to 0.5 according to [8]. Table III presents the scores obtained by the different systems at their best run.

Table III
TREC WEB 2009 DIVERSITY TASK RESULTS

Group Id	Run Id	α-nDCG@10
Waterloo	Uwgym	0.369
uogTr	uogTrDYCcsB	0.282
ICTNET	ICTNETDivR3	0.272
Amsterdam	UamsDancTFb1	0.257

For the *diversity* task, we retained the following runs:

- uwgym: this run acts as a baseline run for the track and should not be considered as an official run. It was generated by submitting the queries to one of the major commercial search engines. The results were filtered to keep only the documents included in the set B of the ClueWeb collection [41];
- uogTrDyCcsB: similarly to the *adhoc* task, this runs relies upon the DPH DFR weighting model but uses a cluster-based query expansion technique, using the Wikipedia documents retrieved [52];
- ICTNETDivR3: this run applies the k-means clustering algorithm to the set of documents retrieved at the *adhoc* task. A document is assigned to the nearest cluster using Euclidean distance or Cosine measure. Each cluster identified represents a subtopic of the query [23];
- UamsDancTFb1: this run uses a sliding window approach that intends to maximize the similarity with the query and, at the same time, to minimize the similarity with the previous selected document. The documents are selected depending on two metrics: Term Filter (TF) and Link Filter (LF). TF focuses on the number of new unique terms to select a new document, while LF uses the new incoming or outgoing links. The document bringing the most new information (links or terms) is selected [26].

2) Results:

a) Overlap of retrieved document sets: As shown in Figure 3, the behavior observed in the previous experiment is more pronounced: the global overlap does not exceed 0.1, even when one hundred retrieved document lists are considered.

These observations are also true when we focus only on relevant and non-relevant documents (see Figure 4), independently of the number of documents considered. In fact, Table IV shows the overlap reaches 0.238 for relevant documents and 0.065 for non-relevant documents when the overall runs (thousand documents) are taken into account. These results confirm our hypothesis that distinct approaches produce distinct results, even if they attempt to reach the same goal.

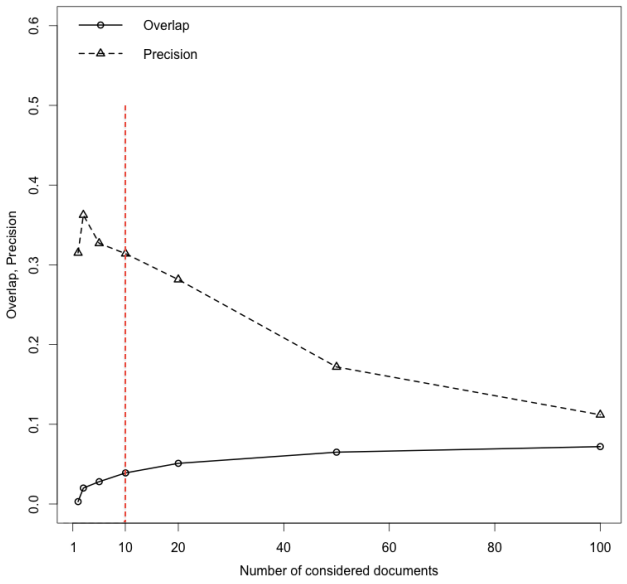


Figure 3. Average global overlap and precision for TREC Web 2009 diversity task

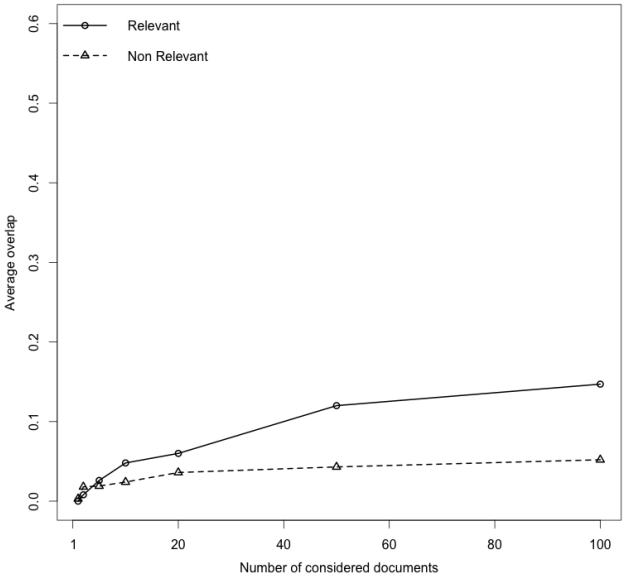


Figure 4. Average overlap for TREC Web 2009 diversity task considering relevant and non relevant documents

b) Aggregating retrieved documents: In the same manner as for the *adhoc* task experiment, we aggregate the runs to check if it helps to bring more *diversity* in the retrieved documents. However we do not use uwgym run which should not be considered as an official run [41]. The aggregation step is also based on the *CombMNZ* function. Finally, we compute α-*nDCG* for the generated run and we obtain 0.283. Although this score stays below the score of the uwgym run which acts as the baseline, it outperforms the best official run (0.282). It confirms that aggregating such approaches produce a more diversified list.

Table IV
GLOBAL OVERLAP FOR THE OVERALL RUNS OF THE TREC WEB 2009 DIVERSITY TASK

Runs	Documents considered	ictnet	uams	uog
uams	Relevant	0.1953		
	Non Relevant	0.0456		
uog	Relevant	0.4051	0.2823	
	Non Relevant	0.1818	0.1052	
uwgym	Relevant	0.1498	0.2095	0.1870
	Non Relevant	0.0154	0.0217	0.0177
Average	Relevant		0.2382	
	Non Relevant		0.0645	

C. Conclusion on the impact of the aggregation method

Whatever is the purpose of the different approaches, whether they intend to diversify the recommended items or whether they are designed to retrieve items matching the users' needs (e.g. topical search), the overlap between the lists of items they retrieve is low. Few documents are retrieved in multiple lists. We note that this observation is especially true when we consider only the first documents, which should theoretically be the most relevant. Finally, the experiment demonstrates that the aggregation of results coming from the selected systems improves accuracy and diversity.

The last experiment we present in Section IV is designed to evaluate the users' perception of diversity and accuracy of a recommendations resulting from the aggregation of various similarity measures. This experiment is conducted thanks to a RS we integrate in a blog platform.

IV. USERS STUDY: THE CASE OF OVERBLOG

A. Diversifying recommendations

We conducted a user experiment to check hypotheses about the relevance of providing diversified recommendations to users in RS while keeping a good level of accuracy. The hypotheses are:

- Most of the time, IR users search for focus information (topicality);
- Sometimes, users want to enlarge the subject they are interested in (topical diversity);
- Some users are in a process of discovering and searching for new information (serendipity);
- The interesting links between documents do not only concern the similarity of their content;
- The integration of diversity in a RS process is valuable because it allows the system to answer additional users' needs.

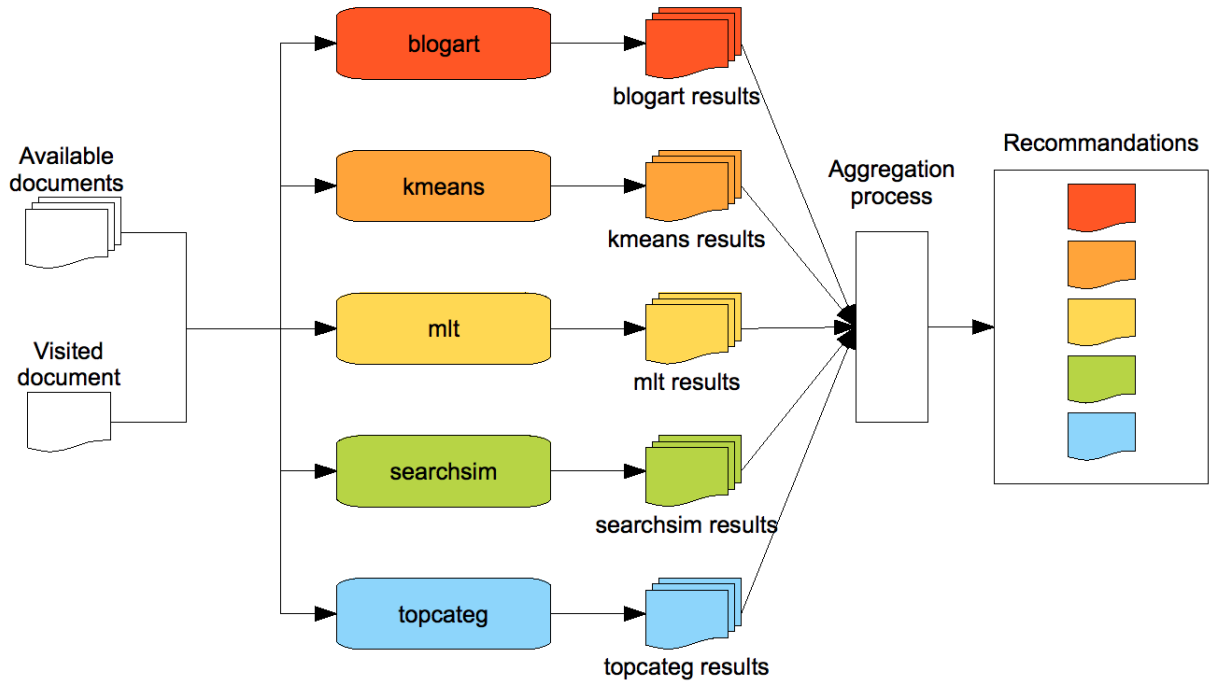


Figure 5. OverBlog aggregation prototype architecture

To check these hypotheses, we recruited 34 Master students in management, fluent in French, and asked them to test and compare various RS. This task lasts about one hour. The users were first asked to type a query on our search engine (first time the query was set, to ensure overlap about the documents they all considered, and then a query of their choice). They had to choose one relevant document and were then shown two lists of recommended documents related to this document:

- One list was based on one of the five similarity measures we designed: *mlt* and *searchsim* that use topicality, *kmeans* that uses topical diversity, and *topcateg* or *blogart* that use serendipity (see system description Section IV-B). These measures act as baselines;
- The other list was our RS, designed by aggregating the results of those five previous defined similarity measures (choosing the first document in the result list for each measure).

Each resulting list contained five documents, and the users did not know which measure it corresponds to. They were then asked to choose which list they found the most relevant, and which one they found the most diversified.

Finally, the two lists were mixed into one, and the users had to assess which documents were relevant according to them.

B. Data and systems

In this experiment, we focused on the French documents of the OverBlog platform [56]. The data used represent more than twenty million articles distributed on more than one million blogs.

We use five similarity measures that have been applied to OverBlog documents to get various recommendation lists which are then aggregated. We define a similarity measure as a function which associates a initial document d_0 (the document visited by the user) with a set of couple (d_i, w_i) where d_i is a document from the available collection and w_i the weight affected to this document by the function. It can be formalized as follows:

$$f(d_0) = \{(d_i, w_i)\} \tag{10}$$

The OverBlog similarity measures are:

- *blogart* (serendipity): returns documents randomly selected in the same blog of the visited document. The author is also the same;
- *kmeans* (topical diversity): classifies the documents retrieved by the Solr search engine [57] with the k-means clustering algorithm [24]. The documents retrieved are those that are most similar to the title of the visited document. We assume each cluster corresponds to one sub-topic of the document subject. The final result list is built by picking up in each cluster the document with the higher score in the initial result list;

- *mlt* (topicality): uses Solr MoreLikeThis module to retrieve similar documents considering all the content of the visited document. The MoreLikeThis module extracts ten representative terms within the visited document. These terms are chosen according to their frequency in the overall corpus and in the document, and then are used to selected similar documents;
- *searchsim* (topicality): uses the Solr search engine which is based on a vector-space model to retrieved documents similar to the title of the visited document;
- *topcateg* (serendipity): retrieves the most popular documents randomly selected in the same category (e.g. "High-tech", "Entertainment", ...) from the OverBlog hierarchy as the visited document. The number of its unique visitors defines the document popularity the day before.

Figure 5 presents the prototype architecture we use to recommend blog articles during the users study. According to this architecture, the available collection and the visited document, each similarity measure independently retrieves an ordered set of documents. These results constitute input data for the aggregation process that picks up the best document from each system. The final recommendation list counts five distinct documents, one per similarity measure.

The use of these five measures aims at simulating the various types of diversity (topicality, topical diversity, serendipity) and intents to limit the overlap between the documents they retrieve.

To ensure that the similarity measures used in the user study retrieve distinct results, we compute the overlap between each pair, similarly to the previous experiments described in Section III.

We observe in Figure 6 the same trends as in the experiments led on the *adhoc* and *diversity* tasks: the overlap is low between the similarity measures based on content similarities (*mlt*, *searchsim* and *kmeans*) and is null in the case of serendipity (*blogart*, *topcateg*).

C. Results

Table V shows the feedback the user panel gave concerning the interest of the proposed lists, and their feeling on the document diversity. For example (4th row), 76.5% of the lists provided by *mlt* measure have been considered as more relevant than the aggregated lists. We can see that the similarity measures perceived as the most relevant are those that focus on topicality.

The aggregated recommendations are seen as more relevant than recommendations coming from other similarity measures roughly once upon two times on average. We get the same result for *blogart* similarity measure. This is more surprising, but confirms that users' expectations sometimes do not concern the document content only.

The answers to the question "Which one of the following result lists seems the most diversified to you?" are even more surprising: there are not high differences between the systems, and the aggregated system is seen, on average, as more diverse in 50% of cases. We think this might be explained by the fact that users have difficulties in defining the notion of diversity. We should have probably helped them by clarifying the question we asked.

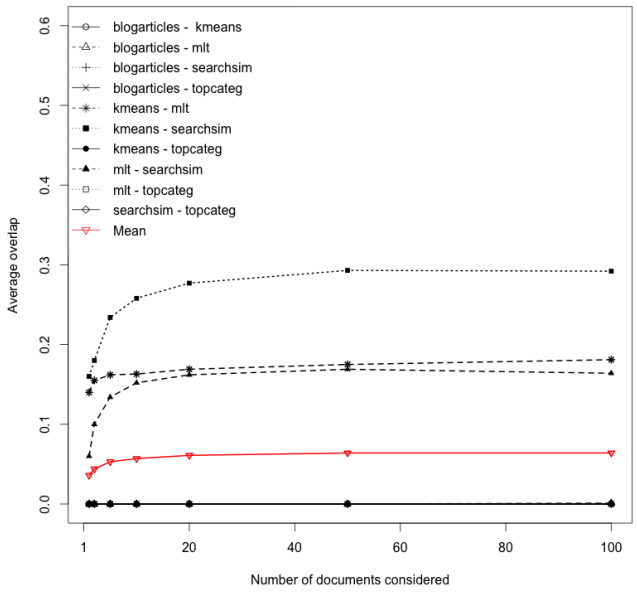


Figure 6. Average overlap for the different results obtained by the five OverBlog similarity

Table V
PERCENTAGE OF USERS WHO CONSIDER THE SYSTEM TO BE MORE RELEVANT/DIVERSIFIED THAN THE AGGREGATED SYSTEM

System	Relevance	Diversity
blogart	44.7%	55.3%
kmeans	70.8%	33.3%
mlt	76.5%	50.0%
searchsim	64.3%	42.9%
topcateg	15.4%	65.4%

Table VI describes the precision of each similarity measure, that is to say the proportion of relevant documents within the retrieved document set. Results confirm the approaches that use content similarities are seen as more relevant. *kmeans*, that proposes topical diversity, has the best results. On the contrary, *topcateg* and *blogart* that search for serendipity have lower results.

As expected, the aggregated recommendations offer an interesting compromise between these different similarity measures and a good balance between diversity (previous result) and precision. Indeed, it obtains a precision value of 0.267 that is higher than the average precision of other similarity measures (0.228). Even if it is lower than the best one (*kmeans*), this result is encouraging regarding the very

Table VI
PRECISION PER SYSTEM

System	*blogart*	*kmeans*	*mlt*	*searchsim*	*topcateg*	*aggregated*
Precision	0.147	0.385	0.265	0.307	0.038	0.267

Table VII
DISTRIBUTION OF THE RELEVANT DOCUMENTS

System: *aggregated* against	*blogart*	*kmeans*	*mlt*	*searchsim*	*topcateg*
Retrieved by the system only	35.00%	52.46%	54.69%	52.43%	8.77%
Retrieved by aggregated system only	65.00%	21.31%	32.81%	38.83%	91.23%
Commons	0.00%	26.23%	12.50%	8.74%	0.00%

low precision value of *topcateg*, *blogart* measures. In fact the low precision value of those measures may introduce noise in the recommendations, which consequently affects the overall precision. At the same time, this loss in precision is not surprising since the result of the aggregation is more diversified: it is considered as more diversified in more than 50% of cases on average. Such negative effect of diversity on accuracy has already been illustrated in [13].

Finally, Table VII compares the aggregated system with the others. It gives the proportion of relevant documents that have been retrieved by each similarity measure. For example, when comparing *mlt* to *aggregated* (4th column), 54.69% of the relevant documents have been retrieved by *mlt* only, 32.81% by *aggregated* only and 12.50% by both only. We can thus observe that, even if more relevant documents come from the similarity measures searching for topicality, a significant part of them comes from the *aggregated* system. Compared to the first experiment (Section III), we think that this result justifies our approach, because more than 20% of relevant documents are retrieved by our system only. It means that one document among the five that are proposed is considered as relevant and would not have been returned when using any system alone.

D. Users study conclusion

The *aggregated* system we propose offers a new framework to combine various similarity measures to recommend items to users. The one implemented and tested here does not outperform the others, but that was not our goal. Rather, our idea is to promote diversity, and we have seen with the user experiments that this is a relevant track. Indeed, by diversifying our recommendations, we are able to answer different and additional users' needs, when the other similarity measures focus on the majority needs: most often the content similarity. The measures we tested for serendipity were quite simple. Nevertheless, the results they returned were considered as relevant by users, and we think this is an encouraging result for improving RS since users are interested in various forms of diversity in result lists.

V. CONCLUSIONS

Users have different expectations when searching and browsing information. Systems that aim at providing tailored results to users should consider this fact. IR systems and RS should aim at answering various facets of the information needs, especially since users become used to be given personalized tools. Diversity in system answers is a way to answer this issue.

In this paper, we have shown the impact of the aggregation of various similarity measures on recommendation diversity.

Our first contribution has been to study the overlap between the documents retrieved by several IR approaches from the literature using the TREC Web 2009 datasets (*adhoc* and *diversity*) and the impact of the aggregation approach on accuracy and diversity. For the *adhoc* task, we have demonstrated that different approaches retrieve different relevant documents even if based on the same aspect of documents as the document topic. The average overlap of the result lists is low, even when the first hundred documents are considered. Moreover, this experiment has underlined an improvement of the accuracy inferred when aggregation is applied. We have also investigated the overlap for topical diversity oriented approaches and obtained similar conclusions: two distinct approaches are unlikely to retrieve the same relevant documents. In the context of topical diversity, we have proved the positive impact of the aggregation approach on recommendation diversity.

Although those approaches are all topical similarity-based, we have noted that they are based on different underlying assumptions, which explains that their overlap is low. The low overlap between the relevant retrieved documents indicates that a perfect system which would be able to satisfy the diversity of the users' needs does not exist, but rather a set of complementary approaches does. This result was the main argument in favor to the approach we defined which aims at aggregating various recommanded item lists.

To validate our proposition in a real context, we conducted a users study. This study aimed at checking if the aggregation of various similarity measures based on topicality

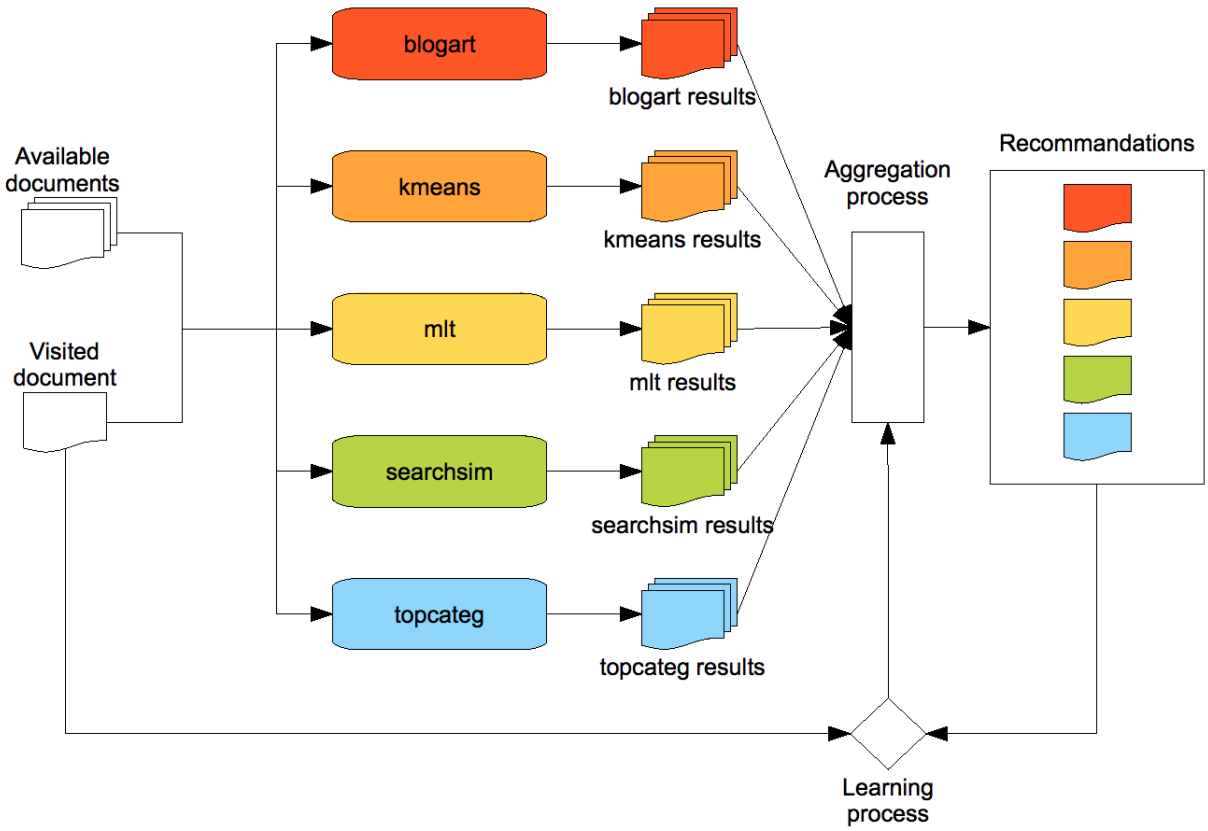


Figure 7. Integration of the learning process in the OverBlog aggregation prototype architecture

(*searchsim*), topical diversity (*kmeans*, *mlt*) or serendipity (*blogart*, *topcateg*) helps to diversify the recommendations and improves the users' satisfaction. We effectively observed a better users' perception of diversity with our RS, without a loss of precision. Indeed, the recommendations resulting from the aggregate similarity measure offer a good balance between accuracy and diversity.

Additionnaly, we promote a framework in which different similarity measures can be combined. One of the main points of this framework is that it is adaptable: any other measure can be added to the framework. It seems that it is worth using approaches that offer serendipity; to this extend, *blogart* seems to be an interesting one. On the other hand, *topcateg* is to be improved.

Our model is not the only one that promote fusing recommandations. Other RS fusion approaches have been proposed in the literature. For example, Schafer *et al.* [58] and Jahrer *et al.* [38] present a "Meta RS". However, when they choose to focus on results shared by the different RS, we instead propose to select the best recommendations from each similarity measure to ensure diversity. We assume that it is important to give a chance to enlarge facets contained in retrieved documents.

When existing approaches focus on designing methods to force diversity in their results (using clustering or MMR), we choose to consider multiple similarity measures to build the recommendation list and ensure diversity. Moreover, it is important that every document may give rise to a wide range of interests for users (a good perception of diversity while keeping a good accuracy level in the recommendation list).

We will direct our future work towards completing the RS architecture to better fit with users' expectations. That is why we will study the learning mechanism to find the proportion of documents coming from every similarity measure, for a given browsed document. As shown on Figure 7, the system may learn the main interests that are important for end-users. To do this, the idea is to use an automatic learning process based on users' feedbacks. We could for example simply initialize the system with equal distribution for each RS (each system contributes equaly to the final list of recommendations), and then increase the proportion of recommendations comming from systems that recommend documents that are more often clicked by the users, and decrease the proportion of recommandation from RS less often considered. Considering the results of the experiments

presented in this paper, we could expect a 80% proportion for topicality systems, and 20% for more original systems. Our future work will also analyze if results are consistent on a real scale experiment using the online blog platform OverBlog when using learning.

REFERENCES

[1] L. Candillier, M. Chevalier, D. Dudognon, and J. Mothe, "Diversity in recommender systems: bridging the gap between users and systems," in *CENTRIC 2011, The Fourth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, 2011, pp. 48–53.

[2] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*. Springer, 2011.

[3] T. Malone, K. Grant, F. Turbak, S. Brobst, and M. Cohen, "Intelligent information-sharing systems," *Communications of the ACM*, vol. 30, no. 5, pp. 390–402, 1987.

[4] M. Montaner, B. López, and J. De La Rosa, "A taxonomy of recommender agents on the internet," *Artificial intelligence review*, vol. 19, no. 4, pp. 285–330, 2003.

[5] A. Kumar and D. Thambidurai, "Collaborative web recommendation systems-a survey approach," *Global Journal of Computer Science and Technology*, vol. 9, no. 5, 2010.

[6] M. Chevalier, T. Dkaki, D. Dudognon, and J. Mothe, "Recommender system based on random walks and text retrieval approaches," in *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases - Discovery Challenge Workshop (ECML/PKDD - DCW)*, T. Smuc, N. Antulov-Fantulin, and M. Morzy, Eds. http://www.irb.hr/: Rudjer Boskovic Institute, 2011, pp. 95–102.

[7] G. Salton and M. McGill, *Introduction to modern information retrieval*. McGraw-Hill, Inc., 1983.

[8] C. Clarke, M. Kolla, G. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon, "Novelty and diversity in information retrieval evaluation," in *31st annual international ACM SIGIR conference on Research and development in information retrieval*, 2008, pp. 659–666.

[9] J. Carbonell and J. Goldstein, "The use of mmr, diversity-based reranking for reordering documents and producing summaries," in *21st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1998, pp. 335–336.

[10] X. Yin, X. Huang, and Z. Li, "Promoting ranking diversity for biomedical information retrieval using wikipedia," *Advances in Information Retrieval*, pp. 495–507, 2010.

[11] A. Chifu and R. T. Ionescu, "Word sense disambiguation to improve precision for ambiguous queries," *Central European Journal of Computer Science*, 2012.

[12] K. Bradley and B. Smyth, "Improving recommendation diversity," in *12th National Conference in Artificial Intelligence and Cognitive Science (AICS-01)*. Citeseer, 2001, pp. 75–84.

[13] G. Adomavicius and Y. Kwon, "Improving aggregate recommendation diversity using ranking-based techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 5, pp. 896–911, 2012.

[14] P. Borlund, "The concept of relevance in ir," *Journal of the American Society for information Science and Technology*, vol. 54, no. 10, pp. 913–925, 2003.

[15] Text REtrieval Conference (TREC) homepage. 04.12.2012. [Online]. Available: http://trec.nist.gov

[16] J. Mothe and G. Sahut, "Is a relevant piece of information a valid one? teaching critical evaluation of online information," *Teaching and Learning in Information Retrieval*, pp. 153–168, 2011.

[17] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *2nd ACM International Conference on Web Search and Data Mining*. ACM, 2009, pp. 5–14.

[18] R. Santos, C. Macdonald, and I. Ounis, "Selectively diversifying web search results," in *19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1179–1188.

[19] Y. Xu and Z. Chen, "Relevance judgment: What do information users consider beyond topicality?" *Journal of the American Society for Information Science and Technology*, vol. 57, no. 7, pp. 961–973, 2006.

[20] F. Radlinski, P. Bennett, B. Carterette, and T. Joachims, "Redundancy, diversity and interdependent document relevance," in *ACM SIGIR Forum*, vol. 43, no. 2. ACM, 2009, pp. 46–52.

[21] C. Ziegler, S. McNee, J. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *14th international conference on World Wide Web*. ACM, 2005, pp. 22–32.

[22] J. He, K. Balog, K. Hofmann, E. Meij, M. Rijke, M. Tsagkias, and W. Weerkamp, "Heuristic ranking and diversification of web documents," DTIC Document, Tech. Rep., 2009.

[23] W. Bi, X. Yu, Y. Liu, F. Guan, Z. Peng, H. Xu, and X. Cheng, "Ictnet at web track 2009 diversity task," DTIC Document, Tech. Rep., 2009.

[24] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 281-297. California, USA, 1967, p. 14.

[25] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.

[26] R. Kaptein, M. Koolen, and J. Kamps, "Result diversity and entity ranking experiments: Anchors, links, text and wikipedia," DTIC Document, Tech. Rep., 2009.

[27] J. Herlocker, J. Konstan, L. Terveen, and J. Riedl, "Evaluating collaborative filtering recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 5–53, 2004.

[28] N. Lathia, S. Hailes, L. Capra, and X. Amatriain, "Temporal diversity in recommender systems," in *33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR10)*, 2010, pp. 210–217.

[29] G. Cabanac, M. Chevalier, C. Chrisment, and C. Julien, "An original usage-based metrics for building a unified view of corporate documents," in *Database and Expert Systems Applications*. Springer, 2007, pp. 202–212.

[30] D. Dudognon, G. Hubert, J. Marco, J. Mothe, B. Ralalason, J. Thomas, A. Reymonet, H. Maurel, M. Mbarki, P. Laublet, and V. Roux, "Dynamic ontology for information retrieval," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*. CID, 2010, pp. 213–215.

[31] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *32nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1994, pp. 133–138.

[32] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734–749, 2005.

[33] I. Esslimani, A. Brun, and A. Boyer, "A collaborative filtering approach combining clustering and navigational based correlations," *Web Information Systems and Technologies*, pp. 364–369, 2009.

[34] L. Jabeur, L. Tamine, and M. Boughanem, "A social model for literature access: Towards a weighted social network of authors," in *Adaptivity, Personalization and Fusion of Heterogeneous Information*. CID, 2010, pp. 32–39.

[35] J. Mothe, C. Chrisment, T. Dkaki, B. Dousset, and S. Karouach, "Combining mining and visualization tools to discover the geographic structure of a domain," *Computers, environment and urban systems*, vol. 30, no. 4, pp. 460–484, 2006.

[36] K. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[37] The Amazon website. 04.12.2012. [Online]. Available: http://www.amazon.com

[38] M. Jahrer, A. Töscher, and R. Legenstein, "Combining predictions for accurate recommender systems," in *16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 693–702.

[39] C. Vogt and G. Cottrell, "Fusion via a linear combination of scores," *Information Retrieval*, vol. 1, no. 3, pp. 151–173, 1999.

[40] M. Chevalier, A. Dattolo, G. Hubert, and E. Pitassi, "Information retrieval and folksonomies together for recommender systems," *E-Commerce and Web Technologies*, pp. 172–183, 2011.

[41] C. Clarke, N. Craswell, and I. Soboroff, "Overview of the trec 2009 web track," DTIC Document, Tech. Rep., 2009.

[42] C. Clarke, N. Craswell, I. Soboroff, and A. Ashkan, "A comparative analysis of cascade measures for novelty and diversity," in *4th ACM international conference on Web search and data mining*. ACM, 2011, pp. 75–84.

[43] C. Hayes, P. Massa, P. Avesani, and P. Cunningham, "An on-line evaluation framework for recommender systems," in *Workshop on Personalization and Recommendation in E-Commerce*. Springer Verlag, 2002.

[44] J. Lee, "Analyses of multiple evidence combination," in *ACM SIGIR Forum*, vol. 31. ACM, 1997, pp. 267–276.

[45] The Lemur toolkit for language modeling and information retrieval. 04.12.2012. [Online]. Available: http://www.lemurproject.org/indri

[46] P. Chandar, A. Kailasam, D. Muppaneni, L. Thota, and B. Carterette, "Ad hoc and diversity retrieval at the university of delaware," in *Text REtrieval Conf*, 2009.

[47] W. Zheng and H. Fang, "Axiomatic approaches to information retrieval-university of delaware at trec 2009 million query and web tracks," DTIC Document, Tech. Rep., 2009.

[48] The Apache Hadoop project homepage. 04.12.2012. [Online]. Available: http://hadoop.apache.org

[49] J. Lin, D. Metzler, T. Elsayed, and L. Wang, "Of ivory and smurfs: Loxodontan mapreduce experiments for web search," DTIC Document, Tech. Rep., 2009.

[50] Terrier IR platform homepage. 04.12.2012. [Online]. Available: http://www.terrier.org

[51] The free encyclopedia Wikipedia homepage. 04.12.2012. [Online]. Available: http://www.wikipedia.org

[52] R. McCreadie, C. Macdonald, I. Ounis, J. Peng, and R. Santos, "University of glasgow at trec 2009: Experiments with terrier," DTIC Document, Tech. Rep., 2009.

[53] A. Spink and B. Jansen, "A study of web search trends," *Webology*, vol. 1, no. 2, p. 4, 2004.

[54] P. Pu, L. Chen, and R. Hu, "Evaluating recommender systems from the user's perspective: survey of the state of the art," *User Modeling and User-Adapted Interaction*, pp. 1–39, 2012.

[55] E. Fox and J. Shaw, "Combination of multiple searches," *NIST Special Publication*, pp. 243–243, 1994.

[56] The OverBlog website. 04.12.2012. [Online]. Available: http://www.over-blog.com

[57] Solr open source enterprise search platform. 04.12.2012. [Online]. Available: http://lucene.apache.org/solr

[58] J. Schafer, J. Konstan, and J. Riedl, "Meta-recommendation systems: user-controlled integration of diverse recommendations," in *11th international conference on Information and knowledge management*. ACM, 2002, pp. 43–51.

The Environment – Application – Adaptation (EAA) Architecture: Introduction and Details of an Open Implementation

Rémi Emonet
Idiap Research Institute
Martigny, Switzerland
remi.emonet@idiap.ch

***Abstract*—This article considers the software problems of reuse and evolution in the context of Ambient Intelligence. The main contribution of the article is the *Environment, Application, Adaptation* (EAA) approach, evolved from state of the art methods used in software engineering and architecture. In the EAA approach, the *applications* are written such that they only reference some abstract functionalities. On the other side, the capabilities of the *environment* are exposed as an individual service. The power of EAA comes from its *adaptation* layer that bridges the gap between capabilities of the environment and functionalities required by the applications. The *adaptation* layer can be dynamically enriched and controlled, giving the end user an easy way to set up the system. The approach is shown to favor development of reusable services and to enable unmodified applications to use originally unknown services. Overall the contributions of the article are: a) the introduction of the EAA approach with an adaptation layer as first-class citizen, b) an illustration through different use cases, c) a feasibility evaluation with implementation details and complete source code available on-line.**

Keywords-Environment; Application; Adaptation; Open Source; Community Architecture; Ambient Intelligence; DCI; SOA; End-User Programming

I. Introduction

With modern devices and technologies, and with sufficient engineering effort, it is relatively easy to implement smart office and smart home applications. Such applications are usually bound to the considered environment and hard to adapt to a new environment. In the context of Ambient Intelligence, such static application design fails because the user is mobile and the environment evolves continuously. Also, an Ambient Intelligence system is always running and is open: new services (of possibly unknown types) are introduced from time to time. The challenge of software architecture for Ambient Intelligence is to provide a way of maximizing reuse and limiting maintenance. For example, applications should not require any modification or redeployment to handle new service types. Our approach tackles this problem and others.

This papers provides additional details over [1], on various aspects of the work. Importantly, many implementation details had been omitted in [1] and were leaving the reader with unanswered interrogations. To improve on this, we provided both more details within the paper and an online release of all the source code necessary to run the experiments, in the form of a "git" repository (see [2]). Together with [1], this article brings the following contributions:

- we review two important software architectures: the Service Oriented Architectures (SOA), which are widely used in Ambient Intelligence and Data Context Interaction (DCI), which is a relatively recent innovation in the design of "traditional" systems and often ignored by the Ambient Intelligence community;
- we combine and adapt SOA and DCI, together with the factory and whiteboard patterns, and propose a new architectural approach that we name Environment, Application, Adaptation (EAA) and that favors reuse and runtime extensibility;
- we illustrate the EAA approach by detailing multiple use cases of applications and showing the advantages of the approach;
- we propose an implementation of the approach using an existing open source service oriented middleware;
- finally, we provide an open-access release of the source code of all the provided use cases to allow introspection, experimentation and reproducibility.

The article is structured as follows: relevant architectural approaches are presented in Section II and we introduce the new EAA architecture in Section III. Section IV introduces the implementation, which is fully detailed with complete examples in Section V. Finally, Section VI provides conclusions and future directions.

II. Related Work and Approach Foundations

Our approach can be seen in continuity with previous architectural concepts. In this section, we introduce the architectural concepts that motivate our approach and we provide discussions about related work.

A. Related Work in Service Oriented Architectures (SOA)

Service Oriented Architectures (SOA) are used in many different contexts ranging from business integration (within and between companies) to Ambient Intelligence. The principle of SOA is to expose software components as "services". Each service encapsulates a particular functionality and provides access to it through a clearly defined interface.

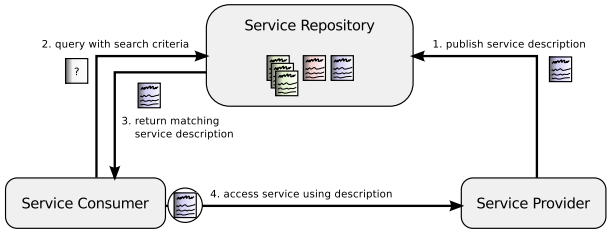


Figure 1. Service discovery, a fundamental aspect of Service Oriented Architectures (SOA). To find concrete providers of the functionality they are looking for, service consumers query a service repository to which all providers are registered. With service discovery, the consumer and the provider are properly decoupled.

One important characteristic of SOA is "service discovery": a service consumer first queries a service repository (or service resolver) to be able to access a matching provider. This discovery process is illustrated in Figure1. Most service oriented frameworks operate with networked services: services that can run on different machines and communicate through a network. A notable exception is OSGi that is broadly used as in [3]. With networked services, one effect of service discovery is to simplify configuration: service consumers only need to know where to find the service repository.

SOA encourages good encapsulation, loose coupling and abstraction. With little effort, it also helps service consumers in reacting to runtime events like the absence or disappearance of a particular service. With encapsulation and discovery, SOA makes it possible to replace a service by another equivalent one, providing the same interface.

As in many other domains, a variety of service oriented initiatives have been proposed but no single standard is clearly dominating. Also, even if service based approaches provide a good way of implementing some "dynamic distributed components", they fail at solving more advanced integration problems.

Consider the use case of having an application dynamically (and with no modification) start using services it was not originally designed to use. Such case is typical of Ambient Intelligence systems where applications and services evolve continuously. SOA allows this if the services have been properly abstracted out and if the integrators make the effort of writing adapter services to bridge the functionality gap. We consider that this integration use case is actually a common one, rather than an exception. Our approach is designed to encourage better abstractions and to make adapter writing a simpler task.

B. Semantic Web Services (SWS) and Service Composition

The convergence of "Semantic Web" and SOA have been trying to solve the integration problem by letting service designers use their own ontology to describe their services. Ontology alignment methods are then used to make correspondences between services from different providers. Using such correspondence, a service for a given provider can be consumed by a consumer that was designed in ignorance of this particular provider.

Multiple approaches mixes web services (WS) technologies with semantic web principles. These are called Semantic Web Services (SWS). Two major set of technologies are used for semantic web services: Web Ontology Language for Services (OWL-S) and the Web Service Modeling Ontology (WSMO) [4]. Both technologies have been very active. An analysis in [5] places WSMO as more promising but less mature than OWL-S; since this analysis was written, WSMO has evolved and matured.

One interesting element of WSMO is the concept of "mediators" that are used to do alignment, conversion or adaptation of different concepts, data and functionalities. From our point of view, this explicit role of mediators is important and close to our approach with an adaptation layer. Depending on the context of use, SWS technologies have some important drawbacks. First, SWS build upon on web service technologies which add complexity and overhead not suitable for certain platforms and developers. Second, service and functionality descriptions in SWS are very detailed, describing IOPE (inputs, outputs, preconditions, effects) of each operation. These details are used to make an automatic, sound and complete reasoning possible, but put an important modeling load on the service writers.

More recently, model driven approaches, such as UWE (UML-based Web Engineering) [6], have been proposed to try to do adaption. However, like other approaches such as [7], the focus is put on the adaptation of graphical user interface to different devices and contexts. The focus is put on proper engineering of web application while ours is to make ambient applications able to evolve and cope with the dynamic nature of the environment. Our proposed approach can actually be seen as orthogonal to such model driven approaches. As web applications are becoming pervasive, both approaches could be put in practice together, replacing our implementation layer by adapting the model driven web engineering technologies.

In Ambient Intelligence, many projects attempt to integrate different services by building upon both SOA and ideas from the semantic web. Fully automatic service composition and adaptation have been explored, e.g., using multi-agent reasoning as in [8]. Some interesting and well designed approaches are [9] and its evolutions. Also, the soft appliances from [10] envision a systematic decomposition of all existing appliances as independent services. In this vision, end-user programming is used to recreate new innovative appliances from services. One of the main difficulty (and limitation) of end-user programming is to make it both accessible to any end user and powerful enough.

As a conclusion, SOA provides a good basis for Ambient Intelligence but it does not ensure good integration capabil-

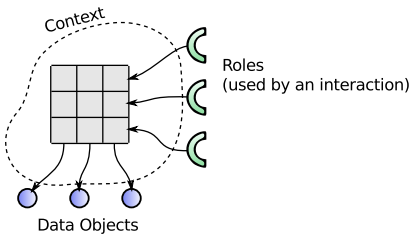


Figure 2. Common representation of the Data, Context, Interaction (DCI) architecture, which complements the Model, View, Controller. Each use case or interaction (I) is implemented using only roles that are fully abstract. The Data (D) are plain objects holding only data and no business specific logic. The Context (C), assembled automatically or through user interaction, is responsible for making some objects play a certain role in the interaction.

ities. Semantic web services are well designed solutions to some of these integration problems but go somewhat to technical and fail at being usable and focused on the adaptation problem. We also think that fully automatic approaches are not desired by the end user: these are not optimal and thus can create frustration, and they prevent end users to express their creativity. Classical end-user programming is also too limited to exhibit, at the same time, these two important aspects: enabling anyone to customize and innovate with applications, and enabling some users to help in integrating new devices. The approach we propose has an explicit adaptation layer and focuses on it, removing the need to describe every possible element in the computational world and making it accessible to most developers.

C. Data Context Interaction: DCI

In our opinion, the most interesting and relevant evolution in recent software architecture and design is the Data Context Interaction (DCI) [11] approach. DCI can be seen as a second attempt to make object orientation (OO) right. The original goal of object oriented programming (and design) was to align the program data model with the user's mental model. This feature is the key to a good human computer interaction: you cannot hide a bad design behind any interface. This becomes more and more important in Ambient Intelligence where user interaction is augmented.

The main principles of DCI are illustrated in Figure2 and can be explained as follows. The *data* objects have the only responsibility to access data (e.g., from a database or memory). In DCI, any use case of the software is a piece of code that manipulates some *roles*, which are fully abstract. A use case is actually an interaction between roles and can be pictured as the scenario involving different roles. A use case uses only a set of roles and never manipulates directly data objects. The concept of *role* together with the *context* are the cornerstone of DCI. A *context* is responsible for doing the mapping of some roles onto some concrete data objects. The context is populated in response to user interaction (e.g., selecting things then clicking on a submit button) and then the use case is executed using this context.

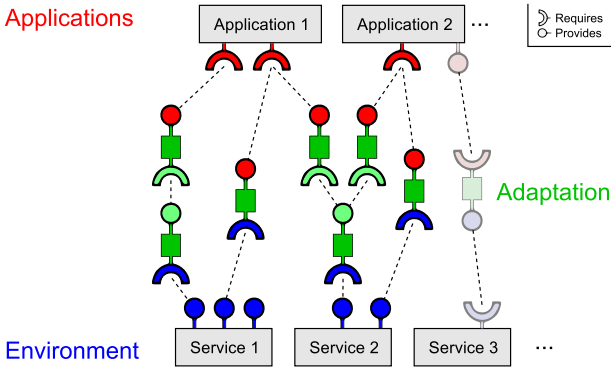


Figure 3. Proposed EAA architecture – *Environment* provides low-level services. *Applications* manipulate only high-level abstract services. *Adaptation* bridges the two and is dynamically extensible and user-controlled. Lighter chain on the right: inversion due to the whiteboard pattern.

As an example, we can consider a banking application with the use case of making a money transfer between two accounts. More precisely, consider the MoneyTransfer use case: it involves three roles that are the SourceAccount role, the DestinationAccount role and the MoneyAmountProvider role. The MoneyTransfer code will start a transaction, then query the amount to transfer from the MoneyAmountProvider, then call $withdraw$ on the SourceAccount and call $credit$ on the DestinationAccount. The context is created and populated by the application when the user is asked to select a source account (e.g., his CheckingAccount data object) and a destination account (e.g., one of his SavingsAccount) and an amount (e.g., could be just a plain "int" value).

D. Other Related Work

A mobile agent is an autonomous program that can migrate between computers over a network. Even if this is an interesting feature for Ambient Intelligence, it can be seen as orthogonal to the subjects discussed in this article and can complement the proposed approach. An example of using mobile agents as an infrastructure is presented in [12].

The domain of human computer interaction tends to evolve from desktop-like applications to Ambient Intelligence. In this context, an emphasis is put on how to dynamically split and distribute user interfaces based on the available devices. The concept of meta-User Interfaces (meta-UI) has been introduced in [7] and consists in having an interface to control and introspect an Ambient Intelligence environment. A deep and interesting analysis related to our problems is conducted in [7], however, their application is limited to the migration and adaptation of graphical user interfaces between devices.

III. Proposed Approach

In this section, we introduce our Environment, Application, Adaptation (EAA) approach and how it can interact with a community built around it. In the same way as DCI

Table I
CORRESPONDENCE BETWEEN DCI AND EAA TERMS, TOGETHER WITH TYPICAL IMPLEMENTATION (OUTSIDE THE WHITEBOARD PATTERN).

DCI Term	EAA Term	Implementation
Data	**E**nvironment	service providers
Context	**A**daptation	adapter factories
Interaction	**A**pplication	service consumers

is an attempt to make OO right (see Section II-C), EAA is an attempt to make SOA right.

A. Environment, Application, Adaptation

The Environment, Application, Adaptation (EAA) approach builds on top of Service Oriented Architectures (SOA) and takes similar inspiration as Data, Context, Interaction (DCI). In EAA, most of the elements are services: in some sense, services act as objects (with interfaces) that can be distributed and dynamically discovered. As in SOA, the capabilities of the *environment* are exposed as plain services in EAA. In a parallel with DCI, these environment services are corresponding to the data part from DCI.

Most importantly, EAA has the equivalent of roles in DCI. Any *application* only manipulates some abstract services (roles) that correspond to its exact requirements. The design of the application is done without bothering about what concrete service can or will be used to fulfill the role. With this choice, the environment will never directly provide any service that an application needs.

In DCI, the context is responsible for the casting: concrete data objects are recruited to play some roles. In EAA, the *adaptation* layer is responsible for the equivalent, which consists in using services from the environment to create services required by the applications. The adaptation layer is populated through implicit or explicit interaction with the end user (same as in DCI).

In Figure 3 (ignoring the lighter rightmost elements), a set of applications, environment services and adapters are shown. Colors are used to distinguish service types coming from the environment (in blue), the applications (in red) or the adaptation (in green). Table I provides a mapping between EAA and DCI terms, and indication on how EAA elements are implemented.

B. Using Service Factories for Adapters

To populate the adaptation layer, some adapter factories are used. Each factory is actually a service that exposes which kind of adapters it can create and that creates them on demand. The concept of service factory is taken from [13] and restricted to adapters: we do not consider the case of "open factories" that can create services without requiring any other service. With our restriction, the number of instantiable adaptation paths becomes finite and it is thus possible to filter and display them to the user (see Section IV).

C. Refinement using the Whiteboard pattern

A useful pattern in service oriented design is the "whiteboard" pattern [14]. The goal of this pattern is to simplify the design of clients of a particular service. Let's consider a Text2Speech service that is designed to receive some text sentence and that outputs it as speech through loud speakers. In a classical approach, any client of the Text2Speech service would first look for the service, then connect to it and then send the message to it. Eventually, the search-and-connect code is here duplicated in all clients.

Using a whiteboard pattern, the situation is reversed and the Text2Speech service is actually doing the search-and-connect. Each client just declares itself as Text2SpeechSource and the Text2Speech will connect to it as soon as it finds it. With the whiteboard pattern, some code is moved from the client to the "server", which limits redundant code writing and makes backward compatible evolutions easier (the server handles the various versions of clients). From a service point of view, now the "server" looks for its clients, which causes an inversion of the provides/requires dependency as shown in Figure 3 (on the right) and in Figure 4 (on the right).

In EAA, the whiteboard pattern is typically used on the view side, i.e., when the application state needs to be brought back to the user (through the environment). The above example of voicing the output of an application using a Text2Speech service is a typical example of this.

D. Community Architecture and Sharing

The structure of the proposed EAA makes it a "community architecture" [15] in a double sense. First, the approach encourages the creation of a community around it and provides a structure for it, and second, it is the community itself that is creating the actual, live, evolving architecture.

We distinguish four entry points in EAA for innovation and extension, each requiring different skills. Compared to some end-user programming approach where there is trade-off to make between the expressive power of the programming and the required skills to use it, EAA has multiple values for this trade-off. It would be interesting to investigate how EAA can be combined with an end-user approach targeting more ease of use than power of expression (higher expression power being provided by EAA).

The first two entry points are for a relatively large audience. First, most end users will be able to innovate at the adaptation level by doing a smart and original choice of adapters for a particular application in their environment. Also, any end user can take part in the community by suggesting new ideas for services, applications or adapters. With proper documentations and examples, we can expect a reasonable part of the users (surely less than 10%) to be able to create new adapters by copying an existing one or using a wizard tool (in current implementation, an adapter

is an XML file that can be easily copied and tuned as shown in following sections).

More advanced extension points concern the contribution of new applications or new environment services. Both require more advanced computer skills but really different ones. Application developers will probably write their application and maybe a couple of adapters to integrate it into the existing ecosystem: the skills required here are mostly classical application development skills. The contributors of new environment services will probably be people that like hacking with new devices or new signal processing methods (image or audio processing, accelerometers, etc.): their goal would be to innovate by providing innovative input or output medium to transform existing applications.

The EAA does not define by itself what kinds of services are used by the people. It is the community itself, by creating new environment services, applications and adapters that decides on what is the actual architecture. We cannot rely on any user to make the best architectural choices. However, if the community is sufficiently large and open, we can expect to find a small proportion of "architects/moderators" as in other open community projects: their role could be for example to avoid proliferation of totally similar concepts and avoid fragmentation of the community.

IV. General Implementation Aspects

To experiment with the proposed approach, we implemented different test cases. The source code of all use cases can be found on-line [2] for additional details and reference. For easier understanding and to allow for reproducibility of the approach, we detail the main aspects of the test case implementation.

Throughout this section and the following, we may reference projects or files from the code available on-line [2]. It is interesting that most of the tasks presented, from coding applications or services to deciding which adapters to use can each be executed by different actors (each with their own skills). This illustrates that most tasks are actually independent and that the resulting system is thus highly extensible.

We implemented the whole presented EAA approach. Most of the tools are written in Java but some commands, usually available under Linux, are used for special functionalities (e.g., text to speech). The "community architecture" aspect, that was totally left out in [1], is now provided. A simple script now allows for an easy download of adapters shared on a central web server. Write access to the server is currently restricted: interested users need to contact us to upload new adapters, or another custom repository can easily be used.

A. The OMiSCID Service Oriented Middleware

Our implementation is based on the open-source OMiSCID [16] service-oriented middleware. A service in OMiSCID can be written in almost any programming language (Java, C++, Python, and most languages running on the JVM) and is discoverable on the network. Each service has a name and may have state variables (also used as properties). In our implementation, we use service variables to expose the information related to the EAA architecture.

Each service can also have connectors. The term "connector" refers to a communication port that can be used to receive messages, broadcast messages or do both. Each message is of arbitrary type but most often either plain text, JSON, or XML. Connectors are the normal mean of passing information between services and we will use it as such.

OMiSCID comes with a graphical user interface (GUI) that can be used to list, monitor and control the services running on a network. The GUI is designed to be highly extensible and allows the user to install plugins in an easy way. As illustrated in the examples and in Figure 4 (detailed later), we designed an interface for the user to decide which adapters should be instantiated among the possible ones. This interface is actually implemented as a plugin for the OMiSCID GUI.

B. Environment Implementation

To compose the environment, we created a set of small reusable functionalities, all exposed as services. Each functionality is actually exposed as a service and an OMiSCID variable "provides" is used to specify which functionalities are provided by the service. In the case of inversion due to the whiteboard pattern, the services from the environment might instead have a "requires" variable.

The developed services that are available online [2] include the following ones: exporting a display area (on a screen or video projector), exporting a mouse pointer, and exporting a "chat" service to allow to open pop-up messages on a computer. Also, under Linux operating systems, we provide additional features like a text-to-speech service based on "espeak", a volume controller and a service to generate synthetic keyboard events on a computer (this one is used for example to control presentations, slide-shows or games). We also provide a computer vision based button to allow user interaction via the real world (e.g., the program detects when the user "clicks" a post-it that he put on a board).

C. Applications Implementation

The applications are also implemented as OMiSCID services that explicitly require some functionalities. The needed functionality is expressed using the "requires" variable. Symmetrically to the environment, when the whiteboard pattern is applied, the "provides" OMiSCID variable is used instead.

The applications that we provide at [2] are a TicTacToe game and a MagicSnake game. It is important to be noted that, thanks to the whiteboard pattern, it is possible to combine multiple services from the environment with only

adapters, without having any application. This can also be seen as a logic-less application. An example of this is to use a button (e.g., any event or a computer vision based button as in Figure 8) to step to the next slide in a presentation.

D. Adaptation Implementation: adapter factories

For the adapters, we designed a generic program that takes an XML description of a family of adapters and starts the corresponding adapter factory (that can start an adapter instance on demand). By convention the service name for adapter factories is AdapterFactory. The XML description contains information about the adapter such as which functionality it takes as "input" and to which one it converts it. The adaptation code, that is usually simple, can be provided within the XML file using languages such as JavaScript, XSLT or dedicated custom languages. Examples of XML descriptions are provided in Section V.

The factory description contains information about what adapter the factory can create (see Figure 7 for an example fully detailed in Section V). A "from" variable contains the name of the functionality that the adapter will take as input. The "to" variable is used for the target functionality that the adapter produces. Some parameters can be used in the "to" variable and are defined in the "parameters" variable. In the "parameters" variable, a special construct can be used to specify that the parameter must take a value that corresponds to an existing requirement (a value present in the "requires" variable of a running service).

E. Adaptation Implementation: user interface

As mentioned in previous section and illustrated in Figure 4, we implemented the control of the adaptation layer as an OMiSCID GUI plugin. This plugin mainly involves Netbeans Platform programming (source to be found in `projects/AdapterFramework`) and won't be detailed here, only its behavior will be described.

Using service discovery, the plugin lists of all relevant services:

- any service having a "provides" variable,
- any service having a "requires" variable,
- any service named AdapterFactory.

Then the plugin displays all provided functionality, together with all required ones and all possible adaptation paths that can get constructed using the running adapter factories. The adapters on the paths are initially not instantiated and displayed using a shaded style. When the user double clicks on an adapter, the GUI plugin automatically formats a message and sends it to the appropriate AdapterFactory that in turn will create the necessary adapter. Once the adapter is started, it changes from shaded to solid in the GUI panel. The user can also easily stop any started adapter, by using a dedicated action provided in the OMiSCID GUI service tree.

V. Detailed Test Cases

To showcase our approach, we detail the case of a simple tic-tac-toe game we developed, starting by a global architecture, which is then detailed.

A. Tic Tac Toe Architecture and Benefits

For now, we consider that the environment contains only two computers, and from each one we export some services: a Display, a Mouse3 (mouse pointer with 3 buttons) and a Text2Speech. In total we thus get six environment services running, three on each computer. Each exported Display service has a unique identifier and follows a whiteboard pattern to connect to any matching DisplaySource it finds. A DisplaySource is expected to send drawing commands to the Display.

The game logic is implemented as a service that exposes a TicTacToeModel functionality. The TicTacToeModel encapsulates the state (current board, current player) and the rules of the game (only the current player can play, who wins, etc.). In addition it also requires some functionalities for the input of the players, more precisely, it needs two Grid3x3Clicker with two different unique identifiers. Following a whiteboard pattern, the game logic automatically connects to the matching Grid3x3Clicker it finds.

To bridge the gap between the environment (Display, Mouse3) and the application (Grid3x3Clicker, TicTacToeModel), we introduced a set of simple adapters. The first ones are for input and can be heavily reused in other context: one adapter converts a three button mouse Mouse3 to a single button mouse Mouse1, the second adapter converts a Mouse1 to a Grid3x3Clicker by converting clicked x, y position to some grid index from 0 to 8. We could have skipped the distinction between Mouse3 and Mouse1 but we kept it as it is useful in some other contexts. On the display side, a specific adapter was written to convert TicTacToeModel to a DisplaySource: the tic-tac-toe state change events are converted to drawing commands such as drawing circles.

By letting the user control the adaptation layer, EAA makes the tic-tac-toe become ambient. The use of properly decoupled services ("SOA done right") makes it possible for the user to dynamically select where and how to display the game and how to control it. EAA, with its explicit adaptation layer, makes it also possible to easily create variations of the game that integrates into an Ambient Intelligence vision. To this end, different adapters can be used. A first adapter, which is simple but specific, transforms the game state (TicTacToeModel) to some short textual output to be processed by a Text2Speech service. A reusable adapter, used for input of the game, could use a SpeechRecognizer and converts voice commands such as "play in three" to a Grid3x3Clicker. In addition to the audio modality, computer vision is also used as a possible input: by sticking post-its

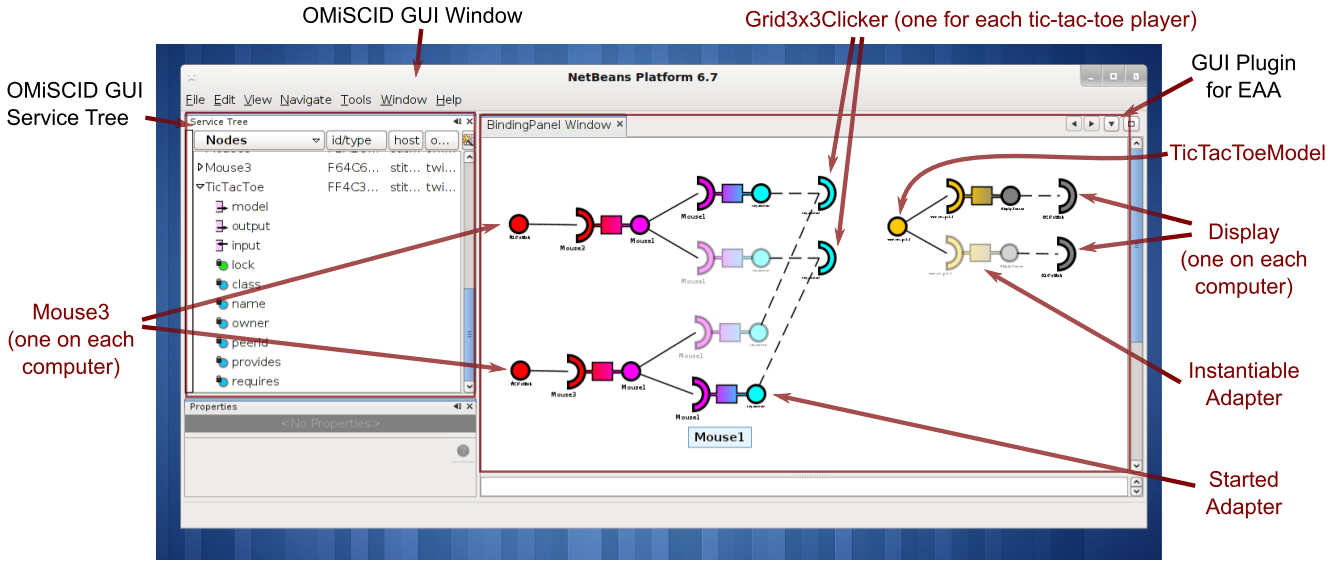


Figure 4. Screen capture of the OMiSCID Graphical User Interface (GUI) for service monitoring and control. On the left side, the default service tree provided by the GUI. On the right, the panel provided by the plugin for the EAA architecture. All provided and required functionalities are shown, using a color for each functionality type. The plugin also considers information from all running adapter factories and proposes all possible adaptation paths to the user. The user can click on a instantiable adapter (shaded), then the plugin queries the corresponding factory for the creation of the adapter. Once it is started, the adapter becomes opaque. Note that due to the whiteboard pattern, the provides/requires relation is reversed for the display side (right part of the panel) where the environment (Display) requires services from the applications (TicTacToeModel).

on a surface, the user can transform it to a Grid3x3Clicker thanks to a dedicated adapter.

B. Tic Tac Toe Implementation Details

The source code for all elements mentioned in this section is provided at [2]. As suggested by the feedback received about [1], we provide a detailed view of how different parts of the system are implemented and articulated. To help in following the explanations of this section, Figure 6 provides a sequence diagram of the interactions between different elements of the system.

Environment – In the provided use case, the environment is populated using a single Java program, the code of which can be found in `projects/ComputerExporter`. The user interface for exporting environment capabilities is shown on the right of Figure 5. This interface can be used to export any number of views, each being a frame, only one being shown on the left of Figure 5. Using multiple views is useful for example when multiple screens or video projectors are plugged to a single computer: in such case, exporting one view per physical display makes more sense.

Each view can be used as input, exporting a "Mouse3" functionality, which is backed by an OMiSCID service, which sends XML messages for each mouse event such as motion events and click events. Each view is also a "Display", also backed by an OMiSCID service that expects to receive messages containing some rendering code fragments. The display service handles multiple simultaneous clients and merges their rendering code fragments. A typical example of a display having multiple clients is the case where we want to display the tic-tac-toe game and also add the rendering of a mouse cursor on top of it (e.g., the cursor of a remote player or a cursor controlled using some hand gestures) as it is the case in Figure 5.

Technically, the rendering code fragments are expressed in JavaScript. The display service sets up a script engine for JavaScript interpretation and fills some context variables so that the snippet can access the rendering context of the frame. Then, the received code fragment is interpreted in context, and this results in graphical elements being drawn in the frame.

Overall, all the services exported by the ComputerExporter are very generic and can be reused over and over for different applications. They are indeed reused in the other use cases presented in Section V-C.

Application – The tic-tac-toe application code can be found in `projects/TicTacToe` and is written in Java. The code simply implements the tic-tac-toe game logic and starts an OMiSCID service. In addition to the variables "provides" and "requires" (presented in previous section), the service has one input connector to receive commands such as "player 1 plays in bin 0", encoded as two digits "10". The service also broadcasts on two output connectors. The connector "model" is used to send the complete game model (expressed in XML) each time a change is made to it. For convenience for the clients, a connector "output" is also used to broadcast events of model changes, i.e., the fact that

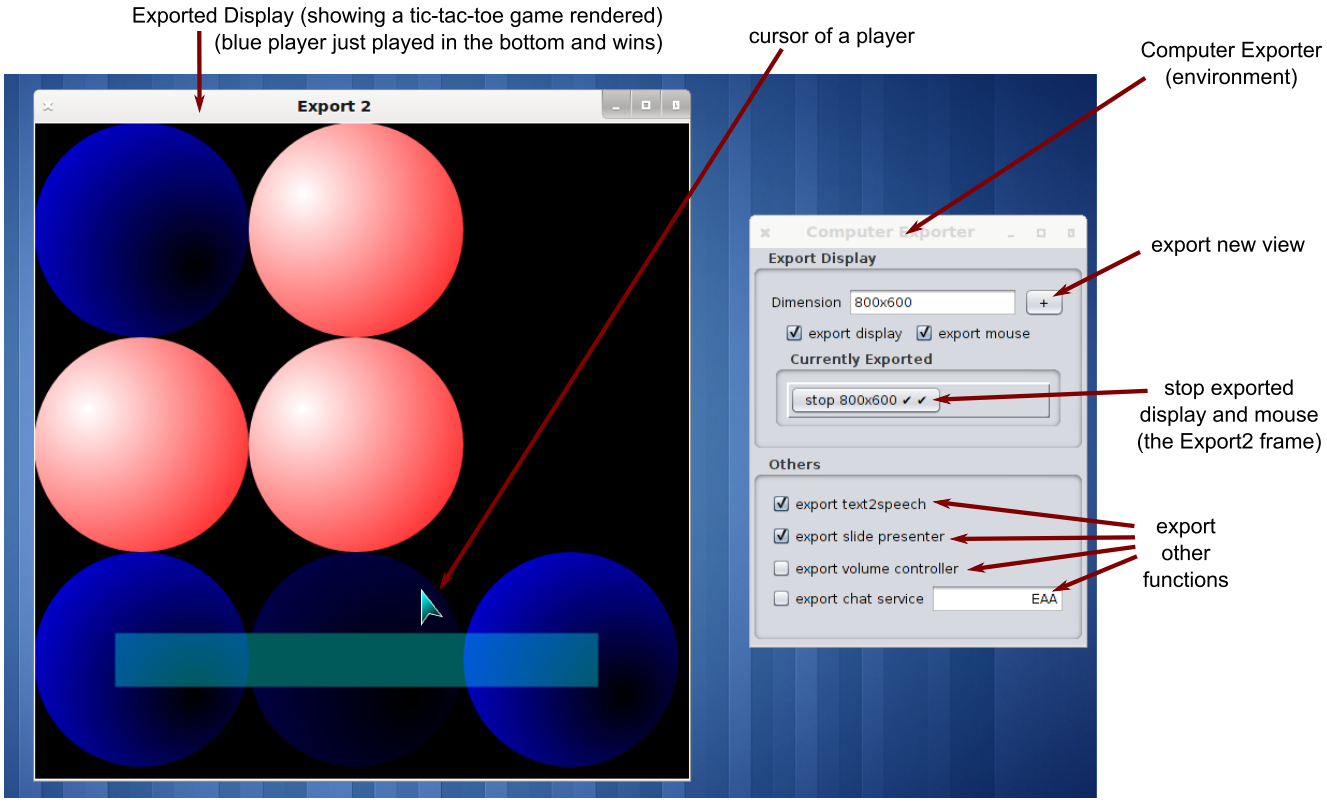


Figure 5. The "ComputerExporter" user interface is shown on the right. It is used to start environment services. On the left is shown an exported display that is currently displaying a tic-tac-toe game. See Section V-B for details.

a given player played at some position, or the fact that one player won. Overall the tic-tac-toe application is minimal, containing only the necessary elements: the game model, the game logic and some interfaces for control and view using an OMiSCID service.

Adaptation – By convention, any OMiSCID service named "AdapterFactory" is considered as a factory of adapters, provided it has the necessary variables introduced in Section IV. Such a factory service has a "create" input connector and when it receives a message on it, it parses the parameters provided inside and starts the corresponding adapter.

Even, if we could write each adapter factory from scratch, we simplified the writing of these. The code, which is generic and shared by all factories, is encapsulated in a java program that can be found in `projects/AdapterTools`. All the information specific to a given adapter factory is included in an XML service description file using conventions that we will illustrate below. For convenience, a script located at `tools/xml-service.sh`, can be called with multiple XML descriptions as parameter and it invokes appropriately the Java program to start all the adapter factories described by the XML files.

Figure 7 illustrates how adapters are written, it provides a complete example of the cursor renderer. The goal might be to create a visual feedback in the same way it is done on classical desktop interfaces to show the mouse pointer position (but here many modalities can be used, e.g., a pointer controlled by hand gestures). As detailed hereafter, this adapters uses the Java2D API through JavaScript and also uses XSLT (Extensible Stylesheet Language Transformations): this adapter can be seen as one of the most complex adapters involved.

The XML service description provided in Figure 7 has a root "service" element and the "name" attribute is used to provide the service name, here "AdapterFactor" on line 1. Then the description contains a succession of OMiSCID variable descriptions: in the example, all these service properties are defined as "constant", meaning they don't change during the lifetime of the service.

Lines 2 to 7 (Figure 7) define the functionality adaptation that this factory can perform. The factory can transform any "Mouse3" functionality, as expressed in the "from" variable. The "to" variable on line 6 is a little more evolved: the factory can create any "Display" source with properties "for" and "z" set according to some parameters. The two variable references "${id}" and "${z}" are references to instantiation parameters, declared in the "parameters" variable detailed

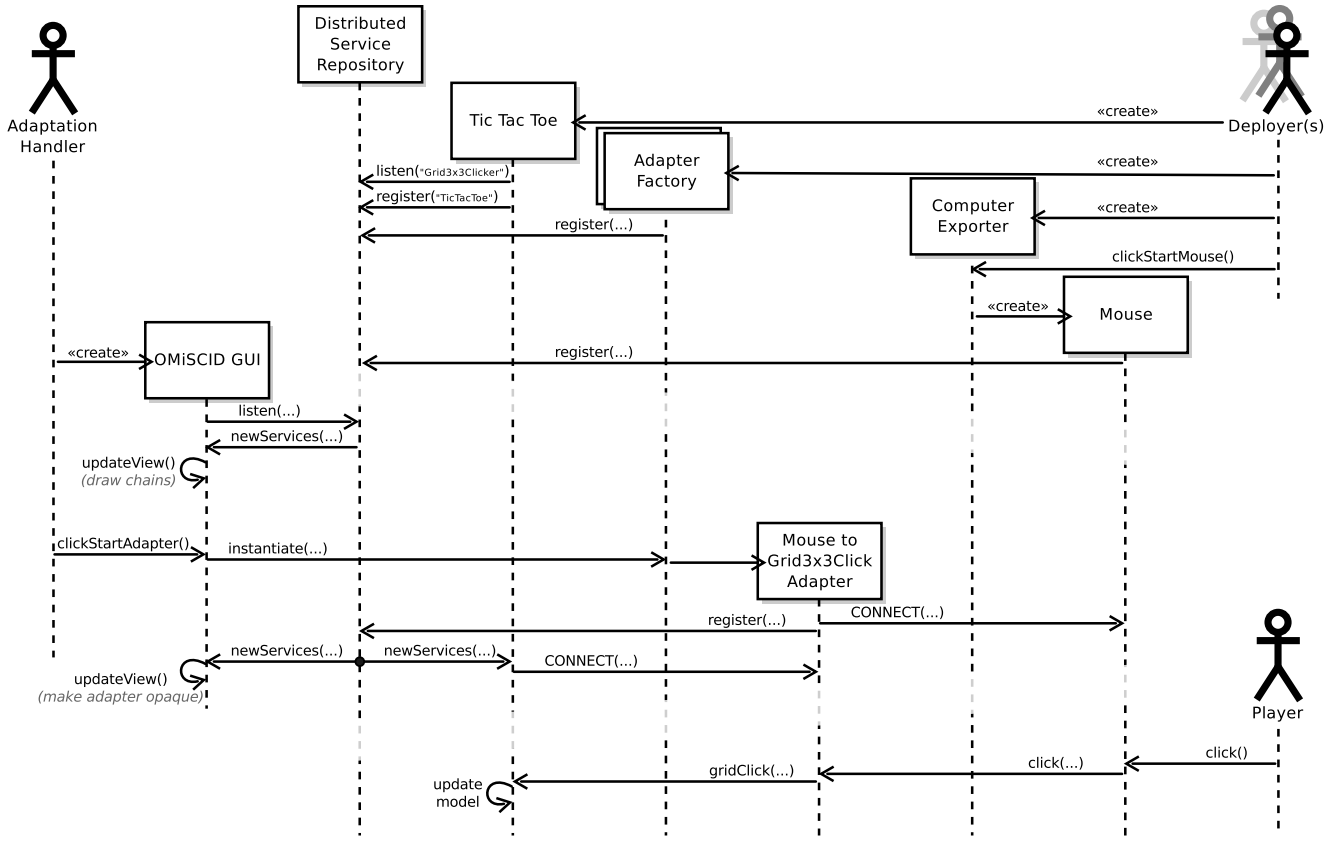


Figure 6. Sequence diagram showing the interactions between different parts of the system. First, the application and some base services are deployed, then the adaptation layer is populated and finally the player plays a turn. Only one input is provided, no display/output for the game is shown in this diagram. See Section V-B for details.

below.

Lines 10 to 14 (Figure 7) define what we call the instantiation parameters. When a client (the main client being the GUI plugin) asks a factory to instantiate an adapter, it sends a message containing the OMiSCID service identifier of the service to adapt (here, the identifier of the Mouse3 to adapt) plus a set of custom instantiation parameters. In the example, there are 5 instantiation parameters. The four last ones are just arbitrary customization parameters, each having a name, a type and a default value. The first parameter in the example, named "id" uses a special construct in place of the default value.

To understand the "#(someRequirement DisplaySource for)" from line 10, we have to remember that the display services from the environment, are actually using a whiteboard pattern. Each Display has a unique identifier. It explicitly "requires" and automatically connects to any service with a "DisplaySource" functionality having a "for" property equal to this unique identifier. The "someRequirement" construct just expresses that the value of the instantiation parameter should correspond to the value of the "for" property of a currently required "DisplaySource" functionality.

Lines 17 to 48 define two other variables that have a special purpose: instead of being solely OMiSCID variables, they also describe the behavior of the adapters. Their respective names "start" and "code" are conventions and these are treated specially by the program that interprets the XML description. The "start" variable describes some (additional) code that will be executed each time an adapter is instantiated by the factory. Starting with "js:", it tells the interpreter that the behavior is expressed in JavaScript. The two lines 19 and 20 respectively add a "display" output connector to the adapter (that it will use to send messages) and create a local input connector that it plugs to the "events" connector of the source service (the Mouse3 to be adapted) on which a listener is registered. With the "listenTo(...)" command, all messages received on the corresponding OMiSCID connector will be processed by the code described in the "code" variable, starting at line 23.

Lines 24 to 44 (Figure 7) define the message handler that each adapter will use to process messages. The code is protected inside an XML CDATA section to avoid escaping all brackets. The code starts with "xslt:" and thus is expressed using the XSLT language which is a standard explicitly

```
<service name="AdapterFactory" ...>
   <variable name="from"> <access>constant</access>
      <value>Mouse3</value>
   </variable>
   <variable name="to"> <access>constant</access>
      <value>DisplaySource for=${id} z=${z}</value>
   </variable>
   <variable name="parameters"> <access>constant</access>
      <value>
         id : string ... = #(someRequirement DisplaySource for)
         size : float ... = 32
         z : float ... = 90
         color1 : Color ... = 0x00FFFF
         color2 : Color ... = 0x000000
      </value>
   </variable>
   <variable name="start"> <access>constant</access>
      <value>js:
         addOutput("display");
         listenTo("events");
      </value>
   </variable>
   <variable name="code"> <access>constant</access>
      <value><![CDATA[xslt:
         <xsl:template match="events/move">
            <message on="display" type="text">
               var x = <xsl:value-of select="@x"/>;
               var y = <xsl:value-of select="@y"/>;
               var s = <xsl:value-of select="$size"/> / 50.0;
               var C = java.awt.Color;
               var c1 = C.decode("<xsl:value-of select="$color1"/>");
               var c2 = C.decode("<xsl:value-of select="$color2"/>");
               g.translate(x, y);
               var p = java.awt.geom.GeneralPath();
               p.moveTo(0,0);
               p.lineTo(s*31.12, s*(50-10.83));
               p.lineTo(s*12.13, s*(50-15.28));
               p.lineTo(0, s*50);
               p.closePath();
               var grad = new java.awt.GradientPaint(s*5, s*5, c1, s*30, s*40, c2);
               g.setPaint(grad);
               g.fill(p);
               g.setColor(C.WHITE);
               g.draw(p);
            </message>
         </xsl:template>
         ]]></value>
      </variable>
</service>
```

Figure 7. XML description of a adapter factory transforming a “Mouse3” functionality into a “DisplaySource” functionality, the goal being to render a cursor on the display at the position of the Mouse3 cursor. See Section V-B for details. For reproduction in this paper, file is reformatted and irrelevant parts are replaced with “...”. In this case the “...” are placeholders for namespace declarations and some type information that is unused in this context.

designed to transform XML documents. The adapters expect to receive XML messages and line 25 defines how the "<move>" messages received on the "events" connector should be processed. Line 26 expresses that (for each move message that is received) the adapter should broadcast a new text message on the "display" OMiSCID connector. Remember that these output messages are eventually reaching a Display service, which will expect some drawing code written in JavaScript. That is exactly what lines 27 to 44 produce: some JavaScript code using an implicit graphic context "g". The JavaScript code is generated by resolving some XSLT variable (using "<xsl:value-of..."). The "@x" and "@y" make use of the XSLT notation to access attributes, here the attributes of the "<move>" element. The "$size", "$color1" and "$color2" use the XSLT notation for accessing variables. These variables actually correspond to the instantiation parameters for the considered adapter and they are brought into the XSLT context by the adapter factory program. In the example, we see that the position from the move message is used to translate the rendered cursor (lines 27, 28, 33), while the parameters (that can be tuned by the client of the factory or by the GUI plugin) control the size and the colors of the cursor.

Figure 9 provides another adapter factory description, transforming a Mouse3 into a Mouse1 by just filtering click events and forwarding move ones. All the concepts involved in this adapter factory description have been covered in previous paragraphs. The main novelty in this adapter lies in the "code" variable at lines 12 to 18. Again the input messages are expected to be XML messages but this time two different possible root elements are handled: the "<click>" and the "<move>" elements on lines 13 and 16. Another difference is that the output messages sent at lines 14 and 17 are not textual messages but rather XML messages (when omitted, the "type" attribute defaults to "xml"). In the case of XML output, the XSLT language is again very well suited as it has been designed for XML transformation.

Figure 10 shows the use of a dedicated language in the "code" variable. This example is actually the one of an adapter from Android key events to some slide controller commands. The language has been designed to make it easy to express a mapping between arbitrary string messages to string messages. XSLT could be used for this purpose but would be unnecessarily verbose. The "code" variable from Figure 10, lines 12 to 15, starts with "map:" indicating the use of the custom mapping language. This language is a domain specific language designed for mapping string to strings. The mappings are given, one by line, each string on the left of the arrow "->" is mapped to the string on the right. In this case, for example, when a message containing "KEY25UP" is received from the "events" connector of the source service, the adapter will broadcast a message containing "next" on its "events" connector. What is not illustrated in this example is the fact that the left part of the arrow is actually a regular expression and the right part a replacement expression. With the "map:" language, most people can create new mappings by copying and updating an existing adapter, without any knowledge of XSLT, JavaScript or even XML.

All adapters are present in a `adapters/` folder in the git repository. The adapters are written using exactly the principles explained previously in this section. Some knowledge of XSLT is required to understand advanced constructs, e.g., as in Figure 7 and Figure 9.

C. Other Test Cases

Apart from the tic-tac-toe game, we also implemented other environment services, applications and adapters.

Games With Analogous Controls – For example we created a MagicSnake game that consists in guiding a snake in a 2D maze to reach a target as fast as possible while avoiding walls. The game can be found in `projects/MagicSnake` and, in the same way as the tic-tac-toe game, it requires a specific controller and exposes its model (but no event-based output). The game is rendered thanks to a dedicated adapter (output illustrated in Figure 8) and reuses the exact same Display service as the tic-tac-toe. As an experiment, we also modified a game called "Nuncabola" where the player controls a ball rolling in a 3D environment. Both games use a two dimensional analog input: we implemented this input with different combinations of environment services and adapters. Eventually, we control these games using:

- obvious device such as a mouse or a keyboard,
- more exotic devices such a accelerometer-based devices (e.g., smart phone, WiiMote) or WiiFit-like devices,
- computer vision and human tracking (e.g., the player moves in the room to control the ball acceleration, or the player moves his hands, arms, etc.)

Various Use of Simple Events – The main use of the "map:" dedicated language that we proposed (see Section V-B) is to easily write adapters for services that exchange only very simple events. We considered various input modalities for such events and various applications and logic-less applications where an adapter links directly an input service from the environment to an actuator service of the environment.

Using simple generation of keyboard events, we implemented a slide presentation controller. We used various methods to skip to the next/previous slide including for example computer vision, e.g., gestures; sound recognition (clapping hands); and voice recognition, e.g., saying "next slide".

We provide an example of computer vision based push button: the code is available in `projects/AdditionalModules` and is actually implemented using a multi-language component framework (see [2]), the assembly of components being defined in

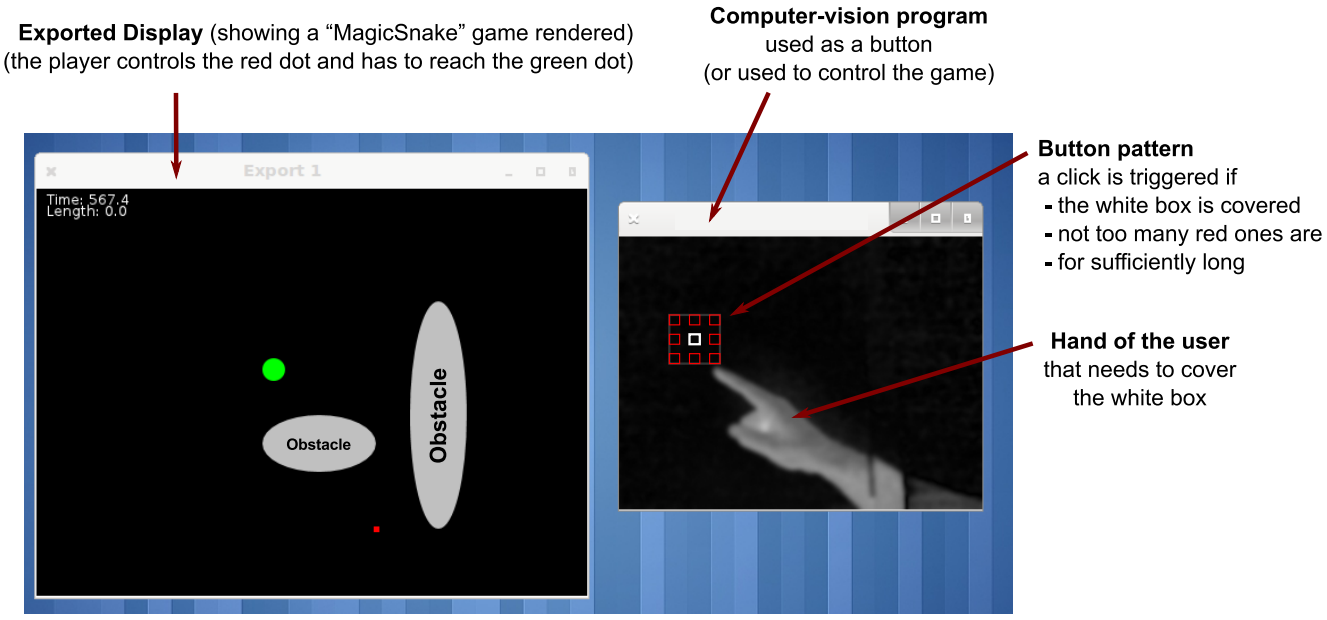


Figure 8. Showing a rendered view from the MagicSnake game together with a debugging window explaining computer vision based environment services. See Section V-C for details. Detecting and tracking the finger tip, its 2D position can be used as an analogue input for the MagicSnake game. We also show a computer-vision based button: we want the button to be triggered if the white box in the center is covered by the user's hand. We also want to avoid unwanted clicks, for example, when the user's hand is fully covering the patterns (including the red boxes) we don't want a click to be triggered.

the file `pipelines/clicklet-omiscid.xml`. The push button is exposed as an OMiSCID service and sends simple "ON" or "OFF" events when it changes state. The overall principle of the vision-based button is to extract the foreground image of the scene, e.g., containing the user or his hand.

Imagining we want a small rectangle (e.g., a post-it) to act as a button, then, we define a region of the size of this rectangle and we detect when it is covered by the foreground pixels. An example of such region is the white box shown in Figure 8. However, in the example of Figure 8, we want a click to be generated if the user covers the white box with his finger. However, if the full hand covers it, then it probably means that the intention of the user is different (reaching another button or just passing between the camera and the post-it). To be able to filter out these wrong clicks, we reason about whether the red boxes from Figure 8 are covered or not (how many of them and which ones). To avoid unwanted repetitive clicks, some delay is added: a button click is validated only if the click detection stays stable for a few frames.

We also implemented a minimal application for Android devices: it exposes an OMiSCID service that sends events each time a physical key is pressed (e.g., volume keys, camera button, etc). The source code is available in `projects/AndroidDeviceExporter`. We could also export a "Display" for the Android device but the Android platform does not implement Java2D and thus it would require to express the rendering in a different way (compared to what we currently used). The Scalable Vector Graphics (SVG) format is a good candidate for an evolution of the display, to have some cross-platform rendering primitives.

Events produced from the Android application can be transformed, for example to a controller for a slide presentation. The corresponding adapter uses the "map:" language and is provided in Figure 10. Similarly, using the computer vision based push button to change slide can be done with a simple mapping like "ON -> next", considering only the press of the button and ignoring when it gets released.

The computer exporter presented in previous sections (source to be found in `projects/ComputerExporter`) can export various services that expect simple string events. The slide presenter accepts four commands: "next" and "previous" for stepping within slides (actually sending left and right arrow key events to the system) and "next+" and "previous+" for skipping directly to the following slide, skipping animations (actually sending up and down arrow key events to the system). There is also a volume controller that accepts "volumeup", "volumedown" and "mute", which impact the system volume in the expected way.

Two services from the computer exporter actually accept any simple string message: the chat service and the text to speech (TTS) service. When it receives a message, the chat service displays the message in a frame on the computer where it is running (the frame is opened if it was closed before). This can be used for popup notifications, application reporting or the textual rendering of models such as the tic-

tac-toe model. Similarly the TTS service will use "espeak" to voice any text message it receives. The TTS can be used for mostly the same purposes as the chat service, the main differences being the throughput (speaking is slower than writing) and the fact that the TTS works even if the user is not facing a screen.

VI. Conclusion and Future Work

This article presented the Environment, Application, Adaptation (EAA) architectural approach. It reuses service oriented principles (SOA) and takes inspiration from the Data, Context, Interaction (DCI) approach. Within our EAA, the environment and the applications are fully independent of each other. This both encourages the design of more generic environment services and eases the deployment of an unmodified application in a new environment: this deployment is possible even if, eventually, the application ends up using only originally unknown services. The glue between what a particular environment offers and what a particular application requires is done by a dedicated adaptation layer. This layer makes the overall system easier to adapt and open to user control and innovation.

An implementation of this approach was showcased: this implementation is fully operational and allows dynamic run-time extension with new services, applications and adapters. To incorporate informal feedback received on [1], we provided very detailed explanations of the mechanisms involved in the implementation. We also cleaned up and made available the source code for all presented use cases on a dedicated page [2].

The foreseen future directions involve the improvement of the user interface (icons for service types, quick filtering, etc), and the exploration of a dedicated interface to create simple adapters based on the mapping language we used. More structured variations of the proposed approach, with different implementation choices, should also be explored: indeed, the approach matches recommendations made by the European IST Advisory Group (ISTAG) in a recent report [17] mentioning that "we might expect to see new programming or modeling languages which include adaptation mechanisms as first-class citizens".

References

[1] R. Emonet, "Environment - Application - Adaptation: a Community Architecture for Ambient Intelligence," in *2011 1st International Conference on Ambient Computing, Applications, Services and Technologies (AMBIENT)*, Oct. 2011.

[2] "Webpage for the source code for the EAA demonstration (this article)." accessed 12-July-2012. [Online]. Available: http://eaa.heeere.com/

[3] C. Escoffier and R. Hall, "Dynamically adaptable applications with iPOJO service components," in *Software Composition*, 2007, pp. 113–128.

[4] ESSI WSMO working group: research and development efforts in the areas of Semantic Web Services, 2007, website and working draft: http://www.wsmo.org/ and http://www.wsmo.org/TR/.

[5] R. Lara, D. Roman, A. Polleres, and D. Fensel, "A conceptual comparison of wsmo and owl-s," in *ECOWS 2004*, ser. LNCS, vol. 3250. Springer, 2004, pp. 254–269. [Online]. Available: http://www.springerlink.com/content/p8358uyre5kw3h7h

[6] J. Preciado, M. Linaje, R. Morales-Chaparro, F. Sanchez-Figueroa, G. Zhang, C. Kroiß, and N. Koch, "Designing rich internet applications combining uwe and rux-method," in *Web Engineering, 2008. ICWE'08. Eighth International Conference on*. IEEE, 2008, pp. 148–154.

[7] J. Coutaz, "Meta-user interfaces for ambient spaces," *Task Models and Diagrams for Users Interface Design*, 2007.

[8] M. Vallée, F. Ramparany, and L. Vercouter, "Dynamic service composition in ambient intelligence environments: a multi-agent approach," in *Proceeding of the First European Young Researcher Workshop on Service-Oriented Computing*, Leicester, UK, April 2005.

[9] M. Assad, D. Carmichael, J. Kay, and B. Kummerfeld, "PersonisAD: distributed, active, scrutable model framework for context-aware services," *Pervasive Computing*, 2007.

[10] J. Chin, V. Callaghan, and G. Clarke, "Soft-appliances: A vision for user created networked appliances in digital homes," *Journal of Ambient Intelligence and Smart Environments*, pp. 69–75, 2009.

[11] J. O. Coplien and G. Bjørnvig, *Lean Architecture: for Agile Software Development*. Wiley, 2010.

[12] R. Razavi, K. Mechitov, G. Agha, and J. Perrot, "Ambiance: a mobile agent platform for end-user programmable ambient systems," in *Proceeding of the 2007 conference on Advances in Ambient Intelligence*. IOS Press, 2007, pp. 81–106.

[13] R. Emonet and D. Vaufreydaz, "Usable developer-oriented functionality composition language (ufcl): a proposal for semantic description and dynamic composition of services and service factories," in *Intelligent Environments, 2008 IET 4th International Conference on*. IET, 2008, pp. 1–8.

[14] O. Alliance, "Listener Pattern Considered Harmful: The Whiteboard Pattern, 2nd rev." http://www.osgi.org/wiki/uploads/Links/whiteboard.pdf, 2004, [Online; accessed 15-December-2012].

[15] F. Moatasim, "Practice of community architecture: A case study of zone of opportunity housing co-operative," Ph.D. dissertation, McGill University, 2005.

[16] R. Emonet, D. Vaufreydaz, P. Reignier, and J. Letessier, "O3miscid: an object oriented opensource middleware for service connection, introspection and discovery," in *International Workshop on Services Integration in Pervasive Environments*, 2006.

[17] ISTAG, "Software Technologies: The Missing Key Enabling Technology," http://cordis.europa.eu/fp7/ict/docs/istag-soft-tech-wgreport2012.pdf, 2012, [Online; accessed 15-December-2012].

```
<service xmlns="..." name="AdapterFactory">
  <variable name="from">...Mouse3...
  <variable name="to"> ... Mouse1...

  <variable name="parameters"> <access>constant</access>
     <value>button1 : int  = 3</value>
  </variable>

  <variable name="start">... js: addOutput("events"); listenTo("events"); ...

  <variable name="code"> <access>constant</access>
     <value><![CDATA[xslt:
       <xsl:template match="events/click[ @button = $button1 ]">
          <message on="events" type="xml"><click button="1" x="{@x}" y="{@y}"/></message>
       </xsl:template>
       <xsl:template match="events/move">
          <message on="events"><move x="{@x}" y="{@y}"/></message>
       </xsl:template>
     ]]></value>
  </variable>
</service>
```

Figure 9. XML description of an adapter factory transforming a Mouse3 functionality into a Mouse1 functionality by simply filtering click messages and forwarding move messages. The “code” of the service uses an “xslt:” transformation, see Section V for details. For reproduction in this paper, file is reformatted and irrelevant parts are replaced with “...”.

```
<service xmlns="..." name="AdapterFactory">
  <variable name="from"> ... AndroidKeys ...
  <variable name="to"> ... RemoteControl for=${id} ...

  <variable name="parameters"> <access>constant</access>
     <value>id: float ... = #(someRequirement RemoteControl for)</value>
  </variable>

  <variable name="start"> ... js: addOutput("events"); listenTo("events"); ...

  <variable name="code"> <access>constant</access>
     <value>map:
        KEY24UP -> next
        KEY25UP -> previous
        KEY80UP -> next
     </value>
  </variable>
</service>
```

Figure 10. XML description of an adapter factory transforming an AndroidKeys functionality into a RemoteControl functionality. The “code” of the service uses a custom “map:” type, see Section V for details. For reproduction in this paper, file is reformatted and irrelevant parts are replaced with “...”.

Rich Annotation Guided Learning

Xiang Li
Computer Science Department
Queens College, CUNY
New York, USA
jackieiuu729@gmail.com

Heng Ji
Computer Science Department
Queens College and Graduate Center, CUNY
New York, USA
hengjicuny@gmail.com

Faisal Farooq
Innovation Center
Siemens Medical Solutions
Malvern, USA
f.farooq@siemens.com

Hao Li
Computer Science Department
Graduate Center, CUNY
New York, USA
haoli.qc@gmail.com

Wen-Pin Lin
Computer Science Department
Queens College, CUNY
New York, USA
danniellin@gmail.com

Shipeng Yu
Innovation Center
Siemens Medical Solutions
Malvern, USA
shipeng.yu@siemens.com

***Abstract*—Supervised learning methods rely heavily on the quantity and quality of annotations provided by humans. As more natural language processing systems utilize human labeled data, it becomes beneficial to discover some hidden privileged knowledge from human annotators. In a traditional framework, a human annotator and a system are treated as isolated black-boxes. We propose better utilization of the valuable knowledge possessed by human annotators in the system development. This can be achieved by asking annotators to provide "rich annotations" for feature encoding. The rich annotations can come at multiple levels such as highlighting and generalizing contexts, and providing high-level comments. We propose a general framework to exploit such rich annotations from human annotators. This framework is a novel extension of our previous work by adding two more levels of rich annotations and two more systematic case studies. To demonstrate the power, generality and scalability of this approach, we apply the method in four very different applications in various domains: medical concept extraction, name translation, residence slot filling and event modality detection. Since richer annotations come at a higher cost (for example, take more time), we investigated the trade-off between system performance and annotation cost, when adding rich annotations from various levels. Experiments showed that the systems trained from rich annotations can save up to 65% annotation cost in order to obtain the same performance as using basic annotations. Our approach is able to bridge the gap between human annotators and systems in a seamless manner and achieve significant absolute improvement (6% - 15%) over state-of-the-art systems for all of these applications.**

***Keywords*-rich annotation; feature engineering**

I. Introduction

As an inter-disciplinary area, statistical natural language processing (NLP) requires two crucial aspects: (1) good choice of machine learning algorithms; (2) good feature engineering. In particular, (2) significantly affects the performance of systems. Linguistic annotation is a fundamental and crucial step of supervised learning. However, feature engineering remains a challenging task because it encompasses feature design, feature selection, feature induction and studies of feature impact, all of which are very time-consuming, especially when there are a lot of data or errors to analyze. As a result, in a typical feature engineering process, the system developer is only able to select a representative data set as the development set and analyze partial errors. Moreover, annotated corpora are usually prepared by a separate group of human annotators before system development. As a result, almost all of the previous NLP systems only utilized direct manual labels for training, while ignoring the valuable knowledge that human annotators have learned and summarized from corpora preparation. In fact, compared to system developers who normally design features based on partial data analysis, human annotators are usually more knowledgeable because they need to go through the entire data set and restrictively follow annotation guidelines.

To draw a parallel, if we consider an NLP system as a "*student*" while a human annotator as a "*teacher*", then the answers or grades (i.e., basic annotations) are just a small part of the pedagogy. Besides grading, a teacher also provides explanations and insights about why an answer is correct or incorrect, comments about what kind of further knowledge the student can benefit from, and how this can be further generalized. Similarly, besides using a text book, a teacher can also highlight part of the content or compose lecture notes. All of these additional evidences and comments can be considered as "*rich annotations*". When human annotators provide certain labels, they rely on certain rationales for the annotation of each instance. The feature engineering is expected to serve as a surrogate for this implicit knowledge. However, in order for that to be accomplished, large amounts of annotated data and highly specialized features are required which is often not feasible.

On the other hand, besides providing the basic annotations, the expert labellers can explicitly provide their rationales for those annotations, which can reduce the amount of training data and thus the annotation effort needed. The challenge then becomes two-fold:

1) feasibility of encoding this extra information such that the machine learning algorithms can exploit. Where as certain additional annotation can automatically be constructed into features, some others would require a systems developer to manually convert the additional information (such as comments) into features. Thus the burden on the system developer (specifically feature engineer) needs to be optimized such that the manual encoding (if necessary) does not require tremendous amount of effort or expertise.
2) the extra annotation effort involved needs to be acceptable to the annotators and the overall cost of the system. For example, simply highlighting the evidence in contexts would not add any significant burden where as generalizing enough knowledge to suggest what kind of linguistic features might be helpful adds slightly more effort. This calls for a fine compromise between the amount of additional information that can potentially make the system better and avoiding burden on the humans.

In this paper, we propose a new and general Rich Annotation Guided Learning (RAGL) framework in order to fill in the gap between an expert annotator and a feature engineer. As an extension of the comment-guided learning framework proposed in our previous work [1], this new framework aims to enrich features with the guidance of all levels of *rich annotations* from human annotators. We will also evaluate the comparative efficacy, generality and scalability of this framework by conducting case studies on four distinct applications in various domains: medical concept extraction, name translation, slot filling and event modality detection. Empirical studies demonstrate that with about little longer annotation time, we can significantly improve the performance for all tasks. We shall measure the annotation cost on these different domains so that this framework is also scalable. For example, the case study on event modality detection demonstrated that the system trained from rich annotations can save 65% annotation cost in order to obtain the same performance as using basic annotations.

The rest of this paper is structured as follows. Section II describes some related work. Section III presents an overview of our new learning framework incorporating rich annotations from human annotators. Section IV, Section V and Section VI present the detailed algorithms to incorporate rich annotations from various levels and four distinct case studies. Furthermore, Section IV illustrates the advantage of Level 1 on medical concept extraction, and Section V shows the contribution of Level 3 on two case studies, name translation and slot filling. After exploring both Level 1 and Level 3, Section VI applies all three levels to a single task, event modality detection, to compare the performance and investigate a trade-off provided by Level 2. Section VII then concludes the paper and sketches the possible future directions.

II. Related Work

In this section, we describe some related work about rich annotations and the applications.

A. Exploiting Rich Annotations

This paper is an extended version of our conference paper published at IMMM2011 [1]. In [1] we only asked human annotators to write down comments and suggestions that might improve re-scoring system output (Level 3 rich annotations) and provided two case studies. In this paper we extended rich annotations to Level 1, 2 and 3, and conducted systematic study on four case studies.

In some NLP tasks such as information retrieval, it's proven effective to incorporate user feedback to customize or tune a system, such as personalized search (e.g., [2]; [3]). However, such user feedback is not always available. Nevertheless most supervised learning methods rely on the labels by human annotators. Therefore there is great potential to fully utilize the deep knowledge from human annotators. [4] proposed to incorporate more of "*teacher's role*" (i.e., privileged knowledge) into traditional machine learning paradigm. We follow this basic idea and incorporate additional feedback from annotators into system development.

Several recent work has pointed out the problem that human annotators are "underutilized" and incorporated rich annotations into many classification problems [5], [6], [7]. Some other work [8], [9], [10] asked human annotators to label or select features. In this paper we shall generalize all kinds of annotator rationales into multiple levels and conduct a systematic study.

Castro et al. [11] investigated a series of human active learning experiments. Our experiment of using Rich Annotation Guided Learning to speed up human assessment exploited assistance from multiple systems.

Our idea of learning from error corrections is also similar to Transformation-based Error-Driven Learning, which has been successfully applied in many NLP tasks such as part-of-speech tagging [12], chunking [13], word sense disambiguation [14] and semantic role labeling [15]. In these applications the transformation rules are automatically learned based on sentence contexts at each iteration. However, our applications require global knowledge that may be derived from diverse linguistic levels and vary from one system to the other, and thus it's not straightforward to design and encode transformation templates. Therefore, in this paper

we choose a more modest way of exploiting the comments encoded by human annotators.

B. Applications

Information extraction from clinical text has recently received a lot of attention. Significant amount of this work in the literature has focused on areas such as radiology and pathology reports [16], [17]. For instance, Taira et al. [18], [19] have performed research on automatic structuring of radiology reports. More recently, researchers are making progress in the automated classification of clinical free text to code [17], [20] and applying machine learning and natural language processing for text mining in systems like BADGER [21], MedLEE [22] and CLEF [23]. Friedman et al [24] discussed the potential of using NLP techniques in the medical domain, and also provides a comparative overview of the state-of-the-art NLP tools applied to biomedical text. Literature in [24], [25], [26] provided a survey of various approaches to information extraction from biomedical text including named entity tagging and extracting relationship between different entities and between different texts. Of direct relevance is the analysis of doctors dictations by Chapman [26], which identified the seven most common uses of negation in doctors dictations. Some of the drawbacks of these works include: i) based on hard coded rules making them difficult to maintain and adapt [21], [23], ii) tuned for specific tasks (e.g., breast care reports [22] or pathology reports [27]) thus failing to generalize, iii) based on institution-specific styles, rules and guidelines (e.g., [28]). In all fairness, this is partially because high quality, labeled datasets of clinical documents have not been available. This is partly due to privacy laws and partly because they are expensive to create. Thus, learning valuable human (in this case clinical) knowledge during the course of annotation would significantly increase the quality of these systems and reduce the annotation efforts at the same time (given we posit that lesser data will need to be annotated).

Name translation is important well beyond the relative frequency of names in a text: a correctly translated passage, but with the wrong name, may lose most of its value. Most of the previous name translation work combined supervised transliteration approaches with Language Model based re-scoring ([29], [30]). Some recent research used parallel corpora or comparable corpora to re-score name transliterations ([31], [32]) or mine name translation pairs ([33], [34], [35], [36], [37], [38], [39], [40], [41]). However, most of these approaches required large amount of seeds and suffered from information extraction errors, and relied on phonetic similarity, context co-occurrence and document similarity to re-score candidate name translation pairs. In contrast, our approach described in this paper does not require any machine translation or transliteration features. Some recent work explored unsupervised or weakly-supervised name translation mining from large-scale data ([41], [42]) and Infoboxes ([43], [44], [45], [46], [47]). For example, Bouma et al. [44] aligned attributes in Wikipedia Infoboxes based on cross-page links; Ji et al. (2009) described various approaches to automatically mine name translation pairs from aligned phrases (e.g., cross-lingual Wikipedia title links) or aligned sentences (bi-texts). Some other work mined name translations from mono-lingual documents that include foreign language texts. For example, Lin et al. [48] described a parenthesis translation mining method; You et al. [49] applied graph alignment algorithm to obtain name translation pairs based on co-occurrence statistics. But none of these approaches exploited the feedback from human annotators.

There are many other alternative automatic assessment approaches for slot filling. Besides the RTE-KBP validation [50] discussed in the paper, some slot filling systems also conducted filtering and cross-slot reasoning (e.g., [51], [52]) to improve results.

Not many methods were proposed to address the problem of event modality attribute. [53] exploited surface features such as part-of-speech tags to detect event modalities and then applied them to improve event coreference resolution. Recent work by [54] described a statistical model based on annotations from rules and crowdsourcing tools. In the meanwhile, a linguistic corpus called "FactBank" with event semantic attributes has been developed by [55].

III. Rich Annotation Guided Learning

In this section, we present the general framework of incorporating rich human annotations into the learning process.

In Table I, we aim to formalize the mapping of some essential elements in human learning and machine learning for NLP.

In regular annotation interface, a human annotator is only asked to provide the final labels (e.g., 0/F or 1/T in binary settings). We call this as the basic annotation in '**Level 0**'. We can see that among these elements, little study has been conducted on incorporating rich annotations from human annotators. In most cases it was not the obligation for the human annotators to write down their evidence or comments during annotation. In contrast, the human learning scenario involves more interactions. However, we can assume that any annotator is able to verify and comment on his/her judgement. We propose to unleash the powerful knowledge based on rich annotations from human annotators on various deeper levels:

- **Level 1**: Ask an annotator to verify a label by providing surface evidence (e.g., highlighting indicative contexts);
- **Level 2**: Ask an annotator to verify a label by providing deep evidence (e.g., generalizing indicative contexts);
- **Level 3**: Ask an annotator to provide comments about linguistic features or resources that might be helpful for system development.

Table I
SOME ELEMENTS IN HUMAN LEARNING AND MACHINE LEARNING FOR NLP

Human Learning	Machine Learning for NLP	Approaches
student	system	baseline NLP system
teacher/ teaching assistant	human annotator/ human assessor	
textbook/ homework answer key	training data with basic annotations	
graded homework		
lecture notes/ graded homework with comments	**training data with rich annotations**	**Our proposed approach**
erroneous homework set	negative samples/errors	transformation based learning
homework review against lecture	system output with background documents	recognizing textual entailment
group study	pooled system responses	voting, learning-to-rank

Entry-level annotators are capable enough to provide rich annotations from Level 1 and Level 2, but we do need some annotators who have some certain knowledge about the task for Level 3 annotations. Based on this intuition we propose a new Rich Annotation Guided Learning (RAGL) paradigm as shown in Figure 1.

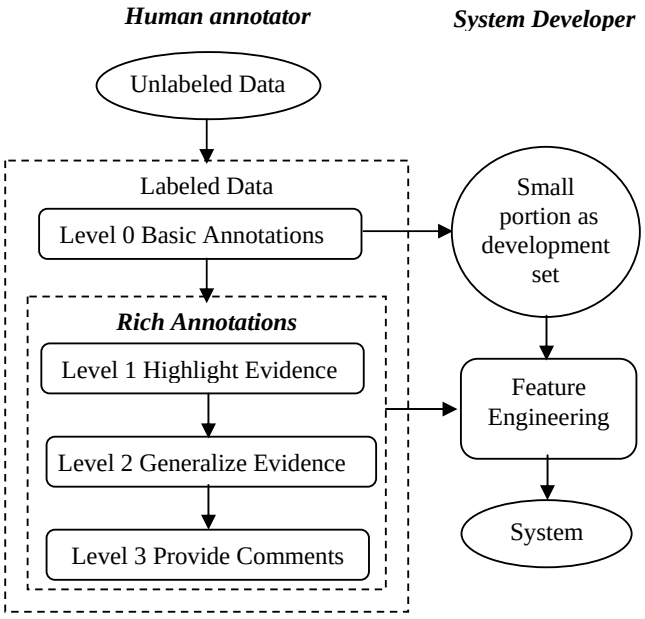


Figure 1. Rich Annotation Guided Learning Framework

IV. LEVEL 1: CHEAP RICH ANNOTATIONS

In this section, we introduce the framework to incorporate rich annotations from level 1, and then apply it to the case study on extracting medical concepts from clinical text.

A. General Framework

While the annotators are providing *basic annotations*, e.g., the class labels, we can often obtain richer annotations at almost no extra cost by highlighting part of text that leads the annotator to that conclusion (often called evidence or rationales). As described earlier in our discussion, this is the Level 1 of rich annotation. For instance, if the text contains the sentence *the patient has no history of alcohol abuse, and does not smoke*, the annotator will label it as *no* for the question whether the patient in question is a smoker or not. In addition, they can also highlight the evidence *does not smoke* since this part of the sentence leads to the no label. Providing this additional evidence adds marginally to the annotation effort, since the annotators would need to read the whole passage regardless, and highlighting the relevant part of the text would be simple if an easy-to-use graphical user interface is provided, such as selecting a contiguous piece of text using the mouse as in Figure 2.

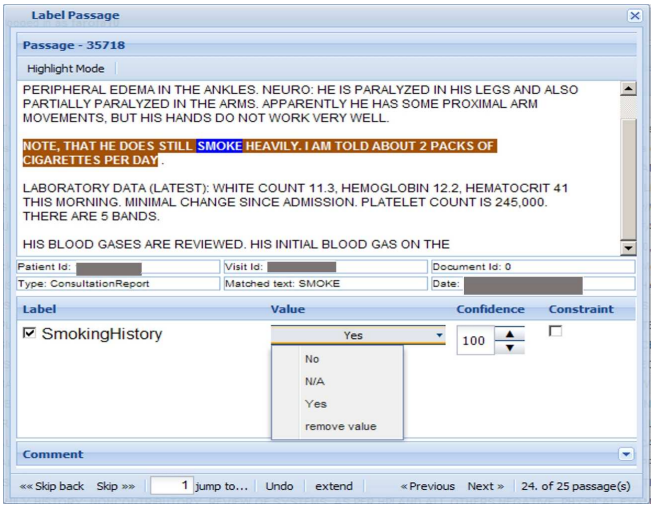


Figure 2. Providing Level 1 Rich Annotations

B. Case Study

For the purposes of this case study, we selected the problem of extracting medical concepts from clinical text. This problem lends itself directly to such a setting because

annotations (provided by clinical experts) are very expensive and often the systems not only have to yield the final answer (Y/N in binary cases), but also the evidence for that answer. Many of these information extraction tasks involve learning of certain medical concepts from the clinical free text, or learning to answer certain clinical questions about the patient. For instance, hospitals in the United States are required by CMS (Centers for Medicare & Medicaid Services) to submit answers of certain quality related questions (called quality measures) after patient discharge. In addition, as part of the meaningful use (MU) of Electronic Medical Records (EMR) initiative under the HITECH Act of the American Recovery and Reinvestment Act, certain key elements need to be reported as well. This is because actionable information that can be regularly and systematically mined from EMRs could lead to improved operational, financial, and clinical outcomes. The answers and the corresponding evidences are often found in the free text medical records (e.g., discharge summary) of the patient. Example questions include *Was the patient given aspirin within 24 hours of admission?*, *Did the adult patient smoke cigarettes anytime during the year prior to hospital arrival?* and *Was the patient assessed for rehabilitation services*. Since these concepts could be represented in different clinical terminologies e.g., rehabilitation assessment can be referred to as physical therapy, occupational therapy, PT, rehab etc., we concentrated on gathering information about five medical concepts with the help of expert medical personnel (such as expert chart abstractors). These concepts were chosen primarily due to their prevalence in quality reporting.

C. Experiments

In our implementation, we choose a state-of-the-art system [56] and show how using rich annotations can significantly improve the classification performance. With this annotation process, providing Level 1 rich annotations does not add a significant burden to the annotators, as they can perform this effortlessly while reading the text. In our experiments, we employed three clinical experts to serve as annotators. Based on past statistics, these annotators spent on average 1.5 hours to annotate 100 examples (of about 200 words each) without providing any rationale. In this new setting, we asked these experts to annotate new datasets with both the class labels as well as highlight rationales by selecting contiguous pieces of text. It was observed that with the rationale, annotation time changed marginally to 1.6 hours for 100 similar such examples, which is $\sim 10\%$ longer. In our experience, the additional annotation effort was acceptable.

Let x_i denote the features for an example text i computed from a dictionary of dimension d. Let $X = x_1, \ldots, x_N$ and $Y = y_1, \ldots, y_N$ denote the N training examples and their labels, respectively. In the training phase, we learn the model parameter $w \in \mathbb{R}^d$ by minimizing a cost function C between X and Y plus a regularization term on w, which can be denoted in the general form as

$$w^* = \arg\min_w \sum_{i=1}^{N} C(y_i, w^T x_i) + \lambda \cdot g(w)$$

with possible constraints on w. Here $g(w) \geqslant 0$ denotes the regularization term on w (such as $\|w\|^2$), and $\lambda \geqslant 0$ is the regularization parameter.

When the rich annotations are available, let $R = r_1, \ldots, r_N$ denote the highlighted evidences, where r_i denote the word sequence of highlighted evidence for example i. The objective is to learn the weight vector w such that the cost function C between X and Y, conditioned on the rich annotations R, is minimized (with regularization). Intuitively, the highlighted evidences provide additional insight to the assigned class label. Since each evidence r_i is simply a sequence of words, let us assume that additional features z_i of dimension s can be induced from the annotations for each example i. With this feature augmentation we can formulate the learning problem as

$$(w^*, v^*) = \arg\min_{w,v} \sum_{i=1}^{N} C(y_i, w^T x_i, v^T z_i) + \lambda_1 \cdot g(w) + \lambda_2 \cdot g(v)$$

with possible constraints on w and v, where $v \in \mathbb{R}^s$ is the weight vector for the evidence-induced feature z_i, and the regularization term involves both w and v (one can also assume a different regularization term for w and v). Then one can use the same solver as the standard binary classification to solve this optimization problem.

Table II
COMPARISON OF LEVEL 0 VS LEVEL 1 ANNOTATIONS FOR MEDICAL CONCEPT EXTRACTION

Medical Concepts	Level 0 AUC	Level 0 Accuracy	Level 1 AUC	Level 1 Accuracy
ST Elevation Assessed	0.87	0.82	0.96	0.90
Assessed for Rehabilitation	0.72	0.74	0.85	0.76
VTE Present on Arrival	0.90	0.83	0.95	0.89
Smoking History	0.72	0.61	0.83	0.80
Joint (e.g., knee) revision	0.87	0.75	0.94	0.80

Experiments were performed using actual EMR data from various medium/large-size hospitals. We built 5 datasets, one for each concept. The questions and the results for the two settings are shown in Table II. These passages were obtained from a set of ~ 10 million sentences by searching, in each case, for a few concepts of interest provided by the clinical experts. For each dataset, we first divided it into two subsets, one held out for testing only (30%) and one used for training (70%).

As the results show, there is significant improvement

across the board in all datasets, both for area under the ROC curve as well as absolute accuracies. This also means that same level of accuracies as Level 0 could have been achieved in the Level 1 setting by annotating fewer data, which would well compensate for the additional effort spent in highlighting.

D. Discussions

Level 1 approach assumes that the annotation evidence exists in the surface texts of the input data, and thus can represent them by simply highlighting such texts at the same time as producing labels. This hypothesis is not valid when a complicated task requires deep understanding of the contexts and external background knowledge. For comparison, we shall present a systematic study on incorporating Level 3 annotations in next section.

V. Level 3: Expensive Rich Annotations

In this section, we present the algorithm to incorporate rich annotations from level 3, and then apply it to the case studies on both name translation mining and slot filling.

A. Algorithm Overview

Recently many NLP tasks have moved from processing hundreds of documents to large-scale or even web-scale data. Once the collection grows beyond a certain size, it is not feasible to prepare a comprehensive answer key in advance. Because of the difficulty in finding information from a large corpus, any manually-prepared key is likely to be quite incomplete. Instead, we can pool the responses from various systems and have human annotators manually review and judge the responses. Assessing pooled system responses as opposed to identifying correct answers from scratch has provided a promising way to generate training data for NLP systems. Usually such tasks require deep knowledge beyond surface information provided by Level 0 and 1. In contrast, the comments from Level 3 can be exploited as features for automatic assessment. Then these features are manually constructed from the comments.

This algorithm aims to extensively incorporate all comments from an old development data set (i.e., "*old homework*" in human learning) into an automatic correction component. This assessor can be applied to improve the results for a new test data set (i.e., "*new homework*" in human learning).

The detailed algorithm can be summarized as follows.

1. The pipeline starts from running the baseline system to generate results. In this step we can also add the outputs from other systems (i.e., classmates in human learning) or even human annotators (i.e., Teaching Assistant (TA) in human learning). We will present one case study on slot filling that incorporates these two additional elements, and the other case study on name translation that only utilizes results from the baseline system.

2. We obtain comments from human annotators on a small development set D^i. Each time we ask a human annotator to pick up N (N=3 in this paper, the value of 3 was arbitrarily chosen; variations in this number of clusters produce only small changes in performance) random results to generate one new comment. One could impose some predefined format or template restrictions for the comments, such as marking the indicative words as rich annotations and encoding them as features. Nonetheless, we found that most of the expert comments are rather implicit and even requires global knowledge. Nonetheless these comments represent general solutions to reduce the common errors from the baseline system.

3. We encode these comments into features through manual construction. We then further train a Maximum Entropy (MaxEnt) based automatic assessor A^i using these features. For each response generated from the baseline system, A^i can classify it as correct or incorrect. We choose a statistical model instead of rules because heuristic rules may overfit a small sample set and highly dependent on the order. In contrast, MaxEnt model has the power of incorporating all comments into a uniform model by assigning weights automatically. In this way we can integrate assessment results tightly with comments during MaxEnt model training.

4. Finally, A^i is applied as a post-processing step to any new data set D^{i+1}, and filter out those results judged as incorrect.

The algorithm can be conducted in an iterative fashion. For example, human annotators can continue to judge and provide comments for D^{i+1} and we can update the automatic assessor to A^{i+1} and apply it to a new data set D^{i+2}, and so on.

We conduct case studies on two distinct application domains: a relatively simple name translation task(V-B) and a more challenging residence slot filling task(V-C).

B. Name Translation Mining

This section presents the first case study of applying Level 3 (human comments) guided learning for name translation validation.

1) Task Definition: Previous name translation pair mining approaches suffer from low accuracy and thus it is important to develop automatic methods to evaluate whether the mined name pairs are correct or not. For example, we need to determine whether the English name "*Michael Jackson*" and the Chinese name "*Mai Ke Er . Jie Ke Xun*" are a correct translation pair. In this paper we focus on validating person name translations by encoding the comments that human annotators made on a small data set.

2) Baseline System: We applied a simple weakly-supervised approach similar to [47] to mine name translation pairs from English and Chinese Wikipedia Infoboxes. A standard Wikipedia entry includes a title, a document describing the entry, and an "infobox", which is a fixed-format

table designed to be added to the top right-hand corner of the article to consistently present a summary of some unifying attributes about the entry. Based on the fact that some certain types of expressions are written in language-independent forms (such as dates and numbers), we generate seed name pairs automatically based on some simple facts (e.g., if two person entries had the same "*birth-date*" and "*death-date*" Infobox slot values, they are considered as a seed pair). Starting from these seeds, we then apply a bootstrapping algorithm based on Infobox slot comparison to mine more pairs iteratively. For example, after we get the seed translation pair of "*Mai Ke Er . Jie Ke Xun (Michael Jackson)*", we can iteratively get new pairs with a large portion of overlapped slots. For example, since "*Ji Xun Wu Ren Zu*" and "*The Jackson 5*" share many slot values such as "*member = Michael Jackson*" and "*years active = 1964-1990*", they are likely to be a translation pair. Next we can use the new pair of "*Ji Xun Wu Ren Zu (The Jackson 5)*" to mine more pairs such as "*Gang Cheng Chang Pian (Steeltown Records)*" their "*labels*".

3) Comments and Feature Encoding: The detailed comments used for validating name translations are as follows.

- **Comment 1: "*these two names do not co-occur often*"** This comment indicates that we can exploit global statistics to filter out some obvious errors, such as "*Ethel Portnoy*" and "*Chen Yao Zu*". Using Yahoo! search engine, we compute the co-occurrence, conditional probability and mutual information of a Chinese Name *CHName* and an English name *ENName* appearing in the same document from web-scale data with setting a threshold for each criteria.
- **Comment 2: "*these two names have very different pronunciations*"** Many foreign names are transliterated from their origin pronunciations. As a result, person name pairs (e.g., "*Lomana LuaLua*" and "*Luo Ma Na . Lu A Lu A*") usually share similar pronunciations. In order to address this comment, we define an additional feature based on the Damerau-Levenshtein edit distance ([57]; [58]) between the Pinyin form of *CHName* and *ENName*. Using this feature we can filter out many incorrect pairs, such as "*Maurice Dupras*" and "*Zhuo Ya . Ke Si Mo Jie Mi Yang Si Qia Ya*".
- **Comment 3: "*these two names have different profiles*"** When human annotators evaluate the name translation pairs, they often exploit their world knowledge. For example, they can quickly judge "*Comerford Walker*" is not a correct translation for "*Sen Gang Er Lang (Jiro Oka Mori)*" because they have different nationalities (one is U.S. while the other is Japan). To address this comment, we define the *profile* of a name as a list of attributes. Besides using all of the Wikipedia Infobox values, we also run a bi-lingual information extraction (IE) system [59] on large comparable corpora (English and Chinese Gigaword corpora) to gather more attributes for *ENName* and *CHName*. For example, since "*Nick Grinde*" is a "*film director*" while "*Yi Wan . Si Te Lan Si Ji*" is a "*physicist*" in these large contexts, we can filter out this incorrect name pair.

The detailed features converted from the above comments are summarized in Table III.

Table III
VALIDATION FEATURES FOR NAME TRANSLATION

Comments	Features
1	co-occurrence, conditional probability and mutual information of *CHName* and *ENName* appearing in the same document from web-scale data
	conditional probability of *CHName* and *ENName* appearing in the same document from web-scale data
	mutual information of *CHName* and *ENName* appearing in the same document from web-scale data
2	Damerau-Levenshtein edit distance between the Pinyin form of *CHName* and *ENName*
3	overlap rate between the attributes of *CHName* and the attributes of *ENName* according to Wikipedia Infoboxes
	overlap rate between the attributes of *CHName* and the attributes of *ENName* according to IE results of large comparable corpora

4) Data and Scoring Metric: We used English and Chinese Wikipedias as of November 2010, including 10,355,225 English pages and 772,826 Chinese pages, and mined 5368 name pairs, where 3719 pairs are correct pairs used as positive samples, and the rest 1649 pairs are incorrect pairs as negative samples. This also shows the capacity of rich annotations to target the problem of unbalanced data. A small set of 100 pairs is taken out as the development set for the human annotator to encode comments. The comment guided assessor is then trained and tested on the remaining pairs by 5-folder cross-validation.

It is time consuming to evaluate the mined name pairs because sometimes the human annotator needs Web access to check the contexts of the pairs, especially when the translations are based on meanings instead of pronunciations. We implemented a baseline of mining name pairs from cross-lingual titles in Wikipedia as an incomplete answer key, and so we only need to ask two human annotators (not system developers) to do manual evaluation on our system generated pairs, which are not in this answer key (1672 in total). A

name pair is judged as correct if both of them are correctly extracted and one is the correct translation of the other. Such a semi-automatic method can speed up evaluation. On average each human annotator spent about 3 hours on evaluation.

5) Overall Performance: Table IV shows Precision (P), Recall (R) and F-measure (F) scores before and after applying the comment guided assessor on name translation pairs. As we can see from Table IV, our approach achieved 28.7% absolute improvement on precision with a small loss (4.9%) in recall. In order to check how robust our approach is, we conducted the Wilcoxon Matched-Pairs Signed-Ranks Test on F-measures. The results show that we can reject the hypothesis that the improvements using Level 3 annotations were random at a 99.8% confidence level.

Table IV
THE IMPACT OF LEVEL 3 ANNOTATIONS ON NAME TRANSLATION

Annotation Type		P (%)	R (%)	F (%)
Basic Annotation	Level 0	69.3	100.0	81.9
Rich Annotation	Level 3	98.0	95.1	**96.5**

C. Slot Filling

In this section, we shall apply Level 3 annotations to a more challenging task of slot filling and investigate the detailed aspects of human comments guided learning by comparing it with other alternative methods.

1) Task Definition: In the slot filling task [60], [61], attributes (or "*slots*") derived from Wikipedia infoboxes are used to create the initial (or reference) knowledge base (KB). A large collection of source news and web documents is then provided to the slot filling systems to expand the KB automatically.

The goal of slot filling is to collect information regarding certain attributes of a query from the corpus. The system must determine from this large corpus the values of specified attributes of the entity. Along with each slot answer, the system must provide the ID of a document that supports the correctness of this answer.

We choose three residence slots for person entities ("*countries_of_residence*", "*stateorprovinces_of_residence*" and "*cities_of_residence*") for our case study because they are one group of the most challenging slot types, for which almost all systems perform poorly (less than 20% F-measure). For example, we need to decide whether it is true that the query "*Adam Senn*" has lived in the country "*America*" or in the city "*Paris*".

2) Baseline Systems: We use a slot filling system [51] that achieved highly competitive results (ranked at top 3 among 31 submissions from 15 teams) at the KBP2010 evaluation as our baseline. This system includes multiple pipelines in two categories: two bottom-up IE based approaches (pattern matching and supervised classification) and a top-down Question Answering (QA) based approach that search for answers constructed from target entities and slot types. The overall system begins with an initial query processing stage where query expansion techniques are used to improve recall. The best answer candidate sets are generated from each of the individual pipelines and are combined in a statistical re-ranker. The resulting answer set, along with confidence values are then processed by a cross-slot reasoning step based on Markov Logic Networks [62], resulting in the final system outputs. In addition, the system also exploited external knowledge bases such as Freebase [63] and Wikipedia text mining for answer validation.

In order to check how robust the RAGL assessor is, we also run it on some other anonymous systems in KBP2010 with representative performance (high, medium and low).

3) Comments and Feature Encoding: The detailed comments used for our slot filling experiment are as follows.

- **Comment 1: "*this answer is not a geo-political name*"**
  This comment is intended to address some obvious errors that could not be Geo-Political (GPE) names in any contexts. In order to address this comment, we apply a very large gazetteer of GPE hierarchy (countries, states and cities) from the geonames website (http://www.geonames.org/statistics/) for answer validation.
- **Comment 2: "*this answer is not supported by this document*"**
  Some answers obtained from Freebase may be incorrect because they are not supported by the source document. Answer validation was mostly conducted on the document basis, but for the residence slots we need to use sentence-level validation. In addition, some sentence segmentation errors occur in web documents. To address this comment, we apply a coreference resolution system [59] to the source document, and check whether any mention of the query entity and any mention of the candidate answer entity appear in the same sentence.
- **Comment 3: "*this answer is not a geo-political name in this sentence*"**
  Some ambiguous answers are not GPE names in certain contexts, such as "*European Union*". To address this comment, we extract the context sentences including the query and answer mentions, and run a name tagger [64] to verify the candidate answer is a GPE name.
- **Comment 4: "*this answer conflicts with this system/other system's output*"**
  When an answer from our system is not consistent with another answer that appears often in the pooled system responses, this comment suggests us to remove our answer. In order to address this comment, we implemented a feature based on hierarchical spatial reasoning. We conduct majority voting on all the available system responses, and collect the answers with global confidence values (voting

weights) into a separate answer set *ha*. Then for any candidate answer *a*, we check the consistency between *a* and any member of *ha* by name coreference resolution and part-whole relation detection based on the gazetteer of GPE hierarchy as described in Comment 1. For example, if "*U.S.*" appears often in *ha* we can infer "*Paris*" is unlikely to be a correct answer for the same query; on the other hand if "*New York*" appears often in *ha* we can confirm "*U.S.*" as a correct answer.

The detailed features converted from the above comments are summarized in Table V.

Table V
VALIDATION FEATURES FOR SLOT FILLING

Comments	Features
1	whether the answer is in the geo-political gazetteer
2	whether any mention of the query entity and any mention of the answer entity appear in the same sentence using coreference resolution
3	whether the answer is a GPE name by running name tagging on the context sentence
4	whether the answer conflicts with the other answers which received high votes across systems using inferences through the GPE hierarchy

4) Data and Scoring Metric: During KBP2010, an initial answer key annotation was created by Linguistic Data Consortium (LDC) through a manual search of the corpus, and then an independent adjudication pass was applied by LDC human annotators to assess these annotations together with pooled system responses to form the final gold-standard answer key. We incorporated the assessment comments for our system output on a separate development set (182 unique non-NIL answers in total) from KBP2010 training data set to train the automatic assessor. Then we conduct blind test on the KBP2010 evaluation data set that includes 1.7 million newswire and web documents. The testing data from our KBP system output consists of 25 positive samples and 121 negative samples, which is also unbalanced. The final answer key for the blind test set includes 81 unique non-NIL answers for 49 queries.

The number of features we can exploit is limited by the unknown restrictions of individual systems. For example, some other systems used distant learning based answer validation and so could not provide specific context sentences. Since comment 2 and comment 3 require context sentences, we trained one assessor using all features and tested it on our own system. Then, we trained another assessor using only comment 1 and 4 and tested it on three other systems representing different levels of performance.

Equivalent answers (such as "*the United States*" and "*USA*") are grouped into equivalence classes. Each system answer is rated as correct, wrong, or redundant (an answer that is equivalent to another answer for the same slot or an entry already in the knowledge base). Given these judgments, we calculate the precision, recall and F-measure of each system, as defined in [60], [61].

5) Overall Performance: Table VI shows the slot filling scores before and after applying the RAGL assessors (because of the KBP Track requirements and policies, we could not mention the specific names of other systems). The Wilcoxon Matched-Pairs Signed-Ranks Test show we can reject the hypothesis that the improvements using RAGL over our system were random at a 99.8% confidence level. It also indicates that the features encoded from comment 2 and comment 3 that require intermediate results such as context sentences helped boost the performance about 3.4%. We can see that although the other high-performing system may have used very different algorithms and resources from ours, our assessor still provided significant gains. Our approach improved the precision on each system (more than 200% relative gains) with some loss in recall. Since most comments focused on improving precision, F-measure gains for moderate-performing and low-performing systems were limited by their recall scores. This is similar to the human learning scenario where students from the same grade can learn more from each other than from different grades. In addition, the errors removed by our approach were distributed equally in newswire (48.9%) and web data (51.1%), which indicates the comments from human annotators reached a good degree of generalization across genres.

Table VI
OVERALL PERFORMANCE OF SLOT FILLING

Slot Filling Systems		Annotation Category	P (%)	R (%)	F (%)
Our system		Level 0	17.1	30.9	22.0
		Level 3 (f1+f4)	26.2	27.2	**26.7**
		Level 3 (full)	38.5	24.7	**30.1**
Other systems	High-Performing	Level 0	13.7	29.6	18.8
		Level 3 (f1+f4)	40.9	22.2	**28.8**
	Moderate-Performing	Level 0	12.2	7.4	9.2
		Level 3 (f1+f4)	35.7	6.2	**10.5**
	Low-Performing	Level 0	6.7	3.7	4.8
		Level 3 (f1+f4)	50	3.7	**6.9**

6) Cost and Contribution of Each Comment: The comments from the RAGL assessor may reflect different aspects of the system. Therefore it will be interesting to investigate what types of comments are most useful and not costly. We did another experiment by applying one comment at a time into the assessor. Table VII shows the results along with the cost of generating and encoding each comment (i.e., knowledge transferring to its corresponding feature), which was carefully recorded by the human annotators.

Table VII
COST AND CONTRIBUTION OF EACH COMMENT

Annotations		Level 0	Level 3			
			f1	f2	f3	f4
Performance	P (%)	17.1	17.6	26.4	26.7	25.6
	R (%)	30.9	30.9	28.4	28.4	27.2
	F (%)	22.0	22.4	27.4	27.5	26.3
Cost	#samples reviewed	-	3	3	3	3
	providing comments (minutes)	-	3	3	3	3
	encoding comments (minutes)	-	30	240	60	30

Table VII indicates that every feature made contributions to precision improvement. Comment 1 (gazetteer-based filtering) only provided limited gains mainly because our own system already extensively used similar gazetteers for answer filtering. This reflects a drawback of our comment generation procedure - the assessor had no prior knowledge about the approaches used in the systems. Comment 2 (using coreference resolution to check sentence occurrence) took most time to encode but also provides significant improvement. Comment 4 (consistency checking against responses with high votes) provided significant gains in precision (8.5%) but also some loss in recall (3.7%). The problem was that systems tend to make similar mistakes, and the human annotator was biased by those correct answers that appeared frequently in the pooled system output. However, Comment 4 was able to filter out many errors that are otherwise very difficult to detect. For example, because "*Najaf*" appears very often as a "*cities_of_residence*" in the pooled system responses, Comment 4 successfully removed six incorrect "*countries_of_residence*" answers for the same query: "*Syrian*", "*Britain*", "*Iranian*", "*North Korea*", "*Saudi Arabia*" and "*United States*". On the other hand, Comment 4 confirmed correct answers such as "*New York*" from "*Brooklyn*", "*Texas*" from "*Dallas*", "*California*" and "*US*" from "*Los Angeles*".

7) Impact of Data Size: We also did a series of runs to examine how our own system performed with different amounts of training data. These experiments are summarized in Figure 3. It clearly shows that the learning curve converges quickly. Therefore, we only need a very small amount of training data (36 samples, 20% of total) in order to obtain similar gains (6.8%) as using the whole training set.

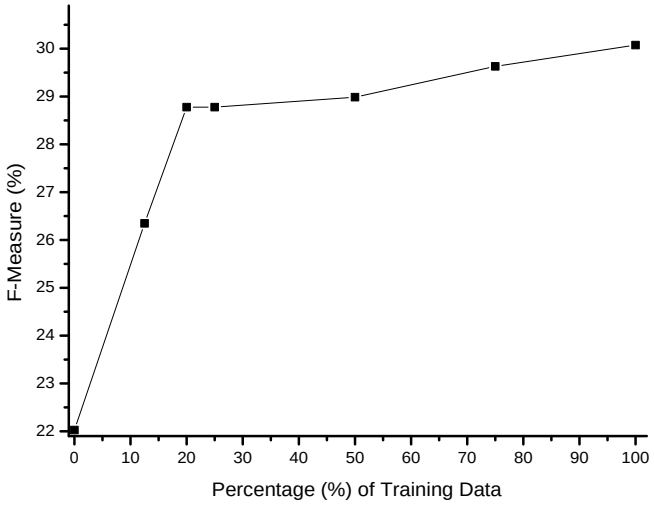


Figure 3. Impact of Training Data Size

8) Speed up Human Assessment: Human assessment for slot filling is also a costly task because it requires the annotators to judge each answer against the associated source document. Since our RAGL approach achieved positive impact on system output, can it be used to as feedback to speed up human assessment? We applied the RAGL assessor trained from comment 1 and comment 4 to the top 13 KBP systems for KBP2010 evaluation set. We automatically ranked the pooled system responses of residence slots according to their confidence values from high to low.

For comparison, we also exploited the following methods:

- **Baseline**
  As a baseline, we ranked the responses according to the alphabetical order of slot type, query ID, query name and answer string and doc ID. This is the same approach used by LDC human annotators for assessing KBP2010 system responses.
- **Oracle (Upper-Bound)**
  We used an oracle (for upper-bound analysis) by always assessing all correct answers first.

Figure 4 summarizes the results from the above 3 approaches. For this figure, we assume a labor cost for assessment proportional to the number of non-NIL items assessed. Note that all redundant answers are also included in these counts because human annotators also spent time on assessing them. This is only approximately correct; it

may be faster (per response) to assess more responses to the same slot. The common end point of curves represents the cost and benefit of assessing all system responses. We can see that if we employ the RAGL assessor and apply some cut-off, the process can be dramatically more efficient than the regular baseline based on alphabetical order. For example, in order to get 79 correct answers (76% of total), RAGL approach took human annotators only 5.5 hours, while the baseline approach took 13.4 hours.

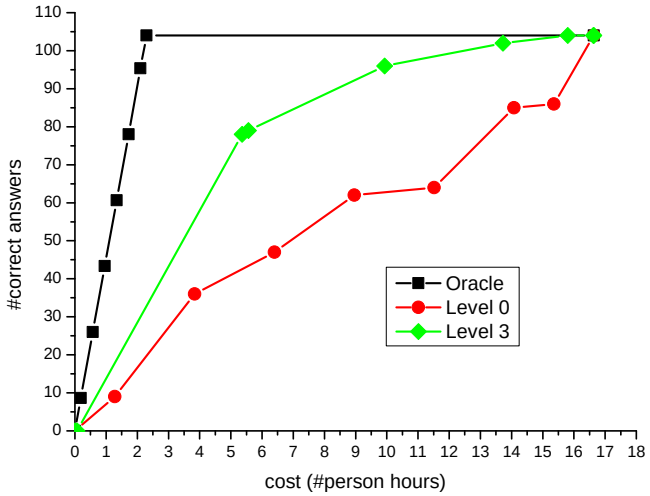


Figure 4. Human Assessment Method Comparison

9) Comparison with Alternative Methods: An alternative approach to validate answers is to use textual entailment techniques as in the RTE-KBP validation task [50], [65], which was partly inspired by CLEF Question Answering task [66]. This task consists of determining whether a candidate answer (hypothesis "*H*") is supported in the associated source document (text "*T*") using entailment techniques. For the residence slots, we are considering in this paper, they treat each context document as a "*T*", and apply pre-defined sentence templates such as "*[Query] lived in [Answer]*" to compose a "*H*" from system output. Entailment and reasoning methods from the TAC-RTE2010 systems are then applied to validate whether "*H*" is true or false according to "*T*". These RTE-KBP systems are limited to individual *H-T* instances and optimized only on a subset of the pooled system responses. As a result, they aggressively filtered many correct answers and did not provide improvement on most slot filling systems (including the representative ones we used for our experiment). In contrast, our RAGL approach has the advantage of exploiting the generalized knowledge and feedback from assessors across all queries and systems.

D. Discussions

We have demonstrated that the comments from Level 3 provided significant improvement for two distinct applications, which require deep understanding of the contexts beyond surface texts. However, we also observed that some comments still require a system developer to fully understand and transfer the knowledge into detailed feature encoding by incorporating external resources. Therefore, the additional cost may vary based on the clarity of each comment and the availability of linguistic resources. In next section we will focus on investigating whether Level 2 is a good trade-off approach between performance gains and cost.

VI. Level 2: A Trade-off

In this section, we will compare three levels of rich annotations by applying all of them to one single task of event modality detection, and investigate whether Level 2 rich annotations can provide a trade-off.

A. Task Definition

We conduct a case study on predicting *Modality* attributes for events defined by the Automatic Content Extraction (ACE) evaluations (http://nist.gov/speech/tests/ace). The Modality attribute indicates whether an event really took place. An event is "Asserted" when the author or speaker makes reference to it as though it were a real occurrence, such as "*A car bomb **exploded** Thursday in the heart of Jerusalem, **killing** at least two people, police said.*". All other events will be annotated as "OTHER", such as "*He asked the committee to **accept** his paper.*" (Commanded and Requested Events), and "*Promises of **aid** made by Arab and European countries.*" (Threatened, Proposed and Discussed Events). The annotators were trained to follow the ACE2005 event annotation guideline: http://projects.ldc.upenn.edu/ace/docs/English-Events-Guidelines_v5.4.3.pdf.

B. General Model

We use a MaxEnt based classifier to detect the modality attribute of a given event instance. This model has the power of assigning weights automatically to features from all levels of rich annotations. During the annotation process, annotators were asked to provide different levels of rich annotations for training data, and we then encoded such rich annotations into a MaxEnt model, as illustrated in the following subsections.

C. Level 0: Baseline

In Level 0, the annotators were asked to label each instance as "Asserted" or "Other" without providing additional markups or comments. During the baseline system development, we selected an n-gram *ng* (n=1, 2, 3) as an indicative context if it matches one of the following two

conditions: (1) *ng* appears only in "Other" events, and with frequency higher than a threshold δ. (2) (the frequency of *ng* occurring in "Other" events) / (the frequency of *ng* occurring in "Asserted" events) is higher than a threshold θ. Both δ and θ were optimized from a small development set including 30 events. The baseline feature is based on the number of indicative context n-grams. For example, given a "*Movement_Transport*" event in "*Bush and Putin were scheduled to leave straight after their talks for the Group of Eight summit of the largest industrialised nations in Evian, France.*" triggered by "*leave*", the indicative context n-grams are "*leave*", "*straight*" and "*to leave*", thus the feature value is 3.

D. Level 1: Highlight Contexts

In Level 1, we ask the human annotators to highlight indicative contexts while labeling modality attribute of each event. The features implemented by the system developers are based on the number of indicative context n-grams, which are highlighted by human annotators. Table VIII presents examples with "Other" modalities and their corresponding highlighted contexts for both Level 1 and Level 2.

E. Level 2: Generalize Contexts

The highlighted features from Level 1 are effective if the contexts are explicit. However, in many other cases the annotators may want to highlight categories of some certain evidence, to indicate informative semantic concepts, templates or patterns that are beyond bag-of-words. For example, instead of highlighting "*scheduled to*", the annotator may want to generalize a category of "*words indicating planning*" because other phrases such as "*plan for*" can play the same role. In Level 2, the human annotators marked up the categories of trigger words and contexts as shown in Table IX, which may indicate "Other" modality. In this Level, annotators only suggest category names, and system developers try to acquire contextual word clusters for each category.

Table IX
CONTEXTUAL CATEGORIES FOR LEVEL 2

Category Name	Size	Example Words/Phrases
Verb	--	(Check whether the event trigger word is a verb)
Modal Auxiliary	10	"could", "would" and "might"
Uncertainty	21	"perhaps", "maybe", "possibly", and "likely"
Planning	11	"planned" and "scheduled"
Assumption	72	"supposed", "estimated" and "expected"
Negation	136	"barely", "impossible", "never", and "declined"

An added advantage of this level of richer annotation is the ease of translation into features. The classification features can, for example, be based simply on the number of matched categories for each event instance.

F. Level 3: Human Comments

Even though Level 2 allows for more flexibility, the annotators are still constrained by existing contexts within the documents. This problem is more concerning in case of sparse databases where the coverage of the explicit contexts is poor. Often, annotators make decisions using global knowledge acquired by aggregating evidence from various resources. This implicit knowledge inferred by annotators cannot be easily represented by highlighted words or categories and thus is captured neither by Level 1 nor Level 2. To address these issues, in Level 3 we allow human annotators to provide verbose comments that represent knowledge about generic situations. In our case study, the expert annotators provided the following comments:

- **Comment 1:** "*If the event is expressed by an entity (person, country, organization, etc.) in a subjective way (e.g., based on assumption, intention, consideration, plans), it's likely to have 'Other' modality. Therefore some contextual libraries including these subjective expressions should be constructed and utilized.*"
- **Comment 2:** "*If the event is likely to occur only under some certain condition, it's likely to have 'Other' modality. Therefore some contextual libraries including these condition expressions should be constructed and utilized.*"

Note that these comments refer to generic guidelines and provide richer knowledge. Some of the comments can be utilized to generate and improve the annotation guideline or train the annotators. However, a barrier in this setting is the translation of these comments into features. In our experiments, system developers manually reviewed and encoded these comments into richer and more generic features. To address these two specific comments, we created *contextual libraries* to cover these broad situations in Table X.

Table X
CONTEXTUAL CATEGORIES FOR LEVEL 3

Category Name	Size	Example Words/Phrases
Expression	116	"expressed", "debated", and "in talks about"
Consideration	77	"like", "consider", and "estimate"
Subjective	77	"assumed", "supposed", and "worried"
Intention	18	"in order to" and "for the purpose of"
Condition	18	"under" and "if"

Consequently, features were generated by checking whether the observed contexts (within the data) include any words/phrases in the above categories.

Table VIII
HIGHLIGHTED CONTEXTS FOR EVENTS WITH OTHER MODALITIES

Event Type	Trigger	Context Sentence with Highlighted Context (in bold/italic)	Expanded Highlighted Context (underlined)
Movement_ Transport	leave	Bush and Putin were ***scheduled to*** leave straight after their talks for the Group of Eight summit of the largest industrialised nations in Evian, France	***scheduled to*** : {plan to, plan for, …}
Conflict_ Attack	attacks	"We are ***warning Israel not to exploit this war*** against Iraq to carry out more attacks against the Palestinian people in the Gaza Strip…	
Justice_ Execute	execute	"***If*** we execute them now we can't bring them to life again should their appeals for a review be granted", said Antasari Azhar…	***If*** : {in case that, with the condition that, whenever, wherever, …}
Life_ Die	death	Indonesia ***will delay*** the execution of six convicts including an Indian on death row after five of them appealed to the Supreme Court for a second review	***will*** : {may, shall, could, would, … } ***delay***
Personnel_ End-Position	leave	Powell, the most moderate member of the Bush cabinet, said he fully agreed with the president's policy on Iraq and ***had no plans to*** leave	
Transaction_ Transfer-Money	payments	***It would be funded*** in two payments of 10 million dollars each upon preliminary and final court approval	***It would*** : {may, shall, could , …} ***be funded***
Transaction_ Transfer-Ownership	sell	The Stalinist state had developed nuclear weapons and ***hinted it may*** sell or use them, depending on US actions.	***hinted it may*** : {shall, could, would, …}
	acquire	Chief executive Andrew Harris said the company ***was likely to abandon plans to*** acquire a hotel in Sydney's Kings Cross red light district…	***was likely*** : {possibly, perhaps, probably, …} ***to abandon plans to***

G. Data and Scoring Metric

We randomly selected a data set from ACE 2005 newswire training set, which consists of 305 "Asserted" events and 305 "Other" events. Because of the data scarcity, ten-fold cross-validation was used to train and test the system.

H. Overall Performance

Table XI shows the accuracy scores when applying features derived from annotations at each level, along with the extra annotation costs compared to basic annotations, which were carefully recorded by the human annotators. The annotation costs do not include the time needed to design annotation scheme or train annotators. We also measured the performance of two human annotators who prepared the ACE 2005 training data on 28 newswire texts (the only subset that includes two-way annotations). As we can see, the system that utilized rich annotations achieved 6.4% absolute improvement over the baseline system using basic annotations, at a 99.9% confidence level according to the Wilcoxon Matched-Pairs Signed- Ranks Significance Test on each folder. We can also see that Level 2 annotation provides more significant gains with small extra cost compared to Level 1, therefore Level 2 can serve as a good trade-off between Level 1 and Level 3.

Figure 5 shows the results when using various size of training data. We can see that rich annotations consistently outperformed basic annotations. It's worth noting that rich annotations using only 25% training data can provide the same accuracy (69.08%) as basic annotations using the full training set. Overall the annotation cost can be reduced from 10 hours to only 3.5 hours.

I. Error Analysis

Analysis on remaining errors show that further advances would require: (1) certain degree of reasoning. For example, although there are many negation context words, the "*Personnel_Elect*" event in the following sentence should still be labeled as "*Asserted*" because the event did happen: "*Of course you will have input into the government but, since you are not directly* ***elected****, it would be a nonsense for you to have direct executive power*". (2) world knowledge such as authority detection of the news source. For example, the following "*Transaction_Transfer-Money*" event should be labeled as "*Other*" because the claims were made by

Table XI
OVERALL PERFORMANCE

Annotation Type		Accuracy	Extra Cost	Extra Effort
Basic Annotation	Level 0	69.02%	0%	––
Rich Annotation	Level 1	71.15%	25%	Highlight contexts
	Level 2	72.95%	5%	Provide category names based on highlighted contexts
	Level 3	70.82%	10%	Provide comments
	Level 1+2+3	75.41%	40%	All of the above

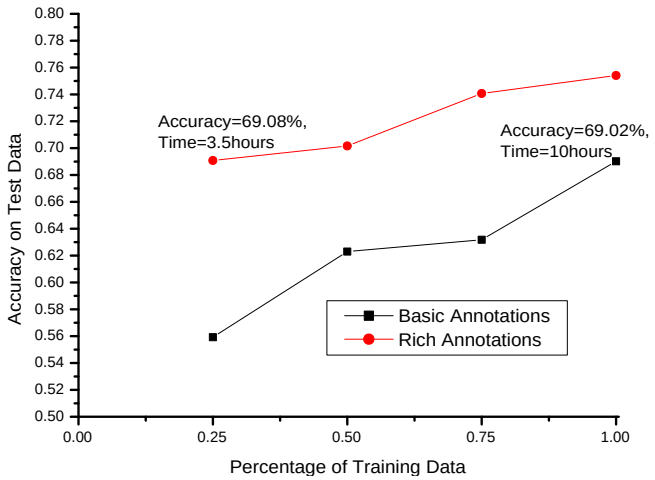


Figure 5. Impact of Training Data Size

an unauthorized source "*the suit*": "*The suit claims Iraqi officials* ***provided*** *money and training to convicted bomber Timothy McVeigh and conspirator Terry Nichols*". These are challenging cases even for human annotators.

VII. CONCLUSION AND FUTURE WORK

In a traditional supervised learning framework, a human annotator and a system are treated as isolated black-boxes to each other. We propose to better utilize the valuable knowledge from human annotators in the system development loop, by asking annotators to provide "rich annotations" for feature encoding. We investigated the trade-off between system performance and annotation cost, when adding rich annotations from various levels. We demonstrated the power and generality of this new framework on four very different case studies. The proposed framework is scalable since we measured the annotation cost on different domains. Experiments showed that the system trained from rich annotations can significantly save annotation cost in order to obtain the same performance as using basic annotations. It also outperformed some traditional validation methods, which, unlike ours, involved a great deal of feature engineering effort. The novelty of our approach lies in its declarative use of the privilege knowledge that human annotators utilize during annotation, which may address some typical errors that a system tends to make. Some of such feedback will be otherwise difficult to acquire for feature encoding (e.g., Comment 3 in name translation and Comment 4 in slot filling). On the other hand, the simplicity of our approach lies in its low cost because it incorporates the bi-product of human annotation, namely their evidence, comments and explanations, instead of tedious instance-based human correction into the learning process. In this way the human annotator's knowledge is naturally transferred to the automatic system. Hence, rich-annotation based learning is amenable to implement but pertinent to a series of common errors identified, and thus fill in the knowledge gap between human annotators and feature engineers.

Remaining error analysis suggested that our future work should focus on mining deeper world knowledge and global reasoning from annotators. Moreover, we will investigate the effects of different rich annotations provided by multiple annotators and also apply on other problem settings. In the future, we are interested in extending this idea to improve other NLP applications and integrating it with human reasoning. Ultimately we intend to investigate automatic ways to prioritize comments and convert comments to features so that we can better simulate the role of teacher in human learning.

VIII. ACKNOWLEDGEMENTS

This work was supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF- 09-2-0053 (NS-CTA), the U.S. NSF CAREER Award under Grant IIS-0953149, the U.S. NSF EAGER Award under Grant No. IIS- 1144111, the U.S. DARPA Broad Operational Language Translations program and PSC-CUNY Research Program. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

REFERENCES

[1] X. Li, W.-P. Lin, and H. Ji, "Comment-guided learning: Bridging the knowledge gap between expert assessor and feature engineer," in *Proc. International Conference on Advances in Information Mining and Management (IMMM2011)*, 2011.

[2] Y. Lv, L. Sun, J. Zhang, J.-Y. Nie, W. Chen, and W. Zhang, "An iterative implicit feedback approach to personalized search," in *Proc. Proc. ACL-COLING2006*, 2006.

[3] S. K. Tyler and J. Teevan, "Large scale query log analysis of re-finding," in *Proc. WSDM2010*, 2010.

[4] V. Vapnik, "Learning with teacher: Learning using hidden information," in *Proc. International Joint Conference on Neural Networks 2009*, 2009.

[5] O. F. Zaidan, J. Eisner, and C. D. Piatko, "Using annotator rationales to improve machine learning for text categorization," in *Proc. NAACL-HLT2007*, 2007.

[6] O. F. Zaidan and J. Eisner, "Modeling annotators: A generative approach to learning from annotator rationales," in *Proc. EMNLP2008*, 2008.

[7] S. Yu, F. Farooq, B. Krishnapuram, and B. Rao, "Leveraging rich annotations to improve learning of medical concepts from clinical free text," in *Proc. ICML 2011 Workshop on Learning from Unstructured Clinical Text*, 2011.

[8] H. Raghavan, O. Madani, and R. Jones, "Active learning with feedback on both features and instances," in *Journal of Machine Learning Research*, 2006.

[9] A. Haghighi and D. Klein, "Prototype-driven learning for sequence models," in *Proc. NAACL-HLT2006*, 2006.

[10] G. Druck, G. Mann, and A. McCallum, "Learning from labeled features using generalized expectation criteria," in *Proc. ACM SIGIR2008*, 2008.

[11] R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, and X. Zhu, "Human active learning," in *Proc. NIPS2008*, 2008.

[12] E. Bril, "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging," in *Computational Linguistics (Volume 21, Number 1, March 1995)*, 1995.

[13] R. L. Milidiu, C. N. dos Santos, and J. C. Duarte, "Phrase chunking using entropy guided transformation learning," in *Proc. ACL-HLT2008*, 2008.

[14] L. Dini, V. D. Tornaso, and F. Segond, "Error driven word sense disambiguation," in *Proc. COLING1998*, 1998.

[15] K. Williams, C. Dozier, and A. McCulloh, "Learning transformation rules for semantic role labeling," in *Proc. CoNLL-2004*, 2004.

[16] D. B. Aronow and K. L. Coltin, "Information technology applications in quality assurance and quality improvement, part ii," *Joint Commission Journal on Quality Improvement*, vol. 10, pp. 465–478, 1993.

[17] Y. Satomura and M. B. do Amaral, "Automated diagnostic indexing by natural language processing," *Medical Informatics*, vol. 3, pp. 149–163, 1992.

[18] D. Johnson, R. Taira, W. Zhou, J. Goldin, and D. Aberle, "Hyperad, augmenting and visualizing free text radiology reports," *RadioGraphics*, vol. 18, pp. 507–515, 1998.

[19] R. Taira, S. Soderland, and R. Jakobovits, "Automatic structuring of radiology free text reports," *RadioGraphics*, vol. 21, pp. 237–245, 2001.

[20] N. Sager, M. Lyman, N. T. Nhan, and L. J. Tick, "Automatic encoding into snomed iii: A preliminary investigation," *Journal of the American Medical Informatics Association*, pp. 230–234, 1994.

[21] S. Soderland, D. Aronow, D. Fisher, J. Aseltine, and W. Lehnert, "Machine learning of text analysis rules for clinical records," *CIIR Technical Report, University of Massachusetts Amherst*, 1995.

[22] X. Zhou, H. Han, I. Chankai, A. Prestrud, and A. Brooks, "Approaches to text mining for clinical medical records," *Proceedings of the 2006 ACM Symposium on Applied Computing (Dijon, France, April 23 - 27, 2006). SAC '06. ACM, New York, NY*, pp. 235–239, 2006.

[23] A. Roberts, R. Gaizauskas, and M. Hepple, "Extracting clinical relationships from patient narratives," *BioNLP 2008: Current Trends in Biomedical Natural Language Processing. Columbus, Ohio, USA.*, pp. 10–18, 2008.

[24] C. Friedman and G. Hripcsak, "Natural language processing and its future in medicine: Can computers make sense out of natural language text," *Academic Medicine*, vol. 74, no. 8, pp. 890–895, 1999.

[25] L. Hirschman, J. C. Park, J. Tsujii, L. Wong, and C. H. Wu, "Accomplishments and challenges in literature data mining for biology," *Bioinformatics*, vol. 18, no. 12, pp. 1553–1561, 2002.

[26] W. Chapman, W. Bridewell, P. Hanbury, G. Cooper, and B. Buchanan, "Evaluation of negation phrases in narrative clinical reports," *Proceedings of the American Medical Informatics Association (AMIA) Symposium*, pp. 105–109, 2001.

[27] A. Coden, G. Savova, I. Sominsky, M. Tanenblatt, J. Masanz, K. Schuler, J. Cooper, W. Guan, and P. C., "Automatically extracting cancer disease characteristics from pathology reports into a disease knowledge representation model," *Biomedical Informatics*, 2009.

[28] G. Savova, K. Kipper-Schuler, J. Buntrock, and C. Chute, "Uima-based clinical information extraction system," *Language Resources and Evaluation: LREC: Towards enhanced interoperability for large HLT systems: UIMA for NLP*, 2008.

[29] Y. Al-Onaizan and K. Knight, "Translating named entities using monolingual and bilingual resources," in *Proc. ACL2002*, 2002.

[30] F. Huang, S. Vogel, and A. Waibel, "Improving named entity translation combining phonetic and semantic similarities," in *Proc. HLT/NAACL2004*, 2004.

[31] R. Sproat, T. Tao, and C. Zhai, "Named entity transliteration with comparable corpora," in *ACL*, 2006.

[32] A. Klementiev and D. Roth, "Named entity transliteration and discovery from multilingual comparable corpora," in *HLT-NAACL*, 2006.

[33] D. Feng, Y. Lv, and M. Zhou, "A new approach for english-chinese named entity alignment," in *Proc. PACLIC*, 2004.

[34] T. Kutsumi, T. Yoshimi, K. Kotani, and I. Sata, "Integrated use of internal and external evidence in the alignment of multi-word named entities," in *Proc. PACLIC*, 2004.

[35] R. Udupa, K. Saravanan, A. Kumaran, and J. Jagarlamudi, "Mint: A method for effective and scalable mining of named entity transliterations from large comparable corpora," in *EACL*, 2009.

[36] H. Ji, "Mining name translations from comparable corpora by creating bilingual information networks," in *ACL-IJCNLP 2009 workshop on Building and Using Comparable Corpora*.

[37] P. Fung and L. Y. Yee, "An ir approach for translating new words from nonparallel and comparable texts," in *COLING-ACL*, 1998.

[38] R. Rapp, "Automatic identification of word translations from unrelated english and german corpora," in *ACL*, 1999.

[39] L. Shao and H. T. Ng, "Mining new word translations from comparable corpora," in *COLING2004*, 2004.

[40] M. Lu and J. Zhao, "Multi-feature Based Chinese-English Named Entity Extraction from Comparable Corpora," in *Proc. PACLIC*, 2006.

[41] A. Hassan, H. Fahmy, and H. Hassan, "Improving Named Entity Translation by Exploiting Comparable and Parallel Corpora," in *Proc. RANLP2007*, 2007.

[42] A. E. Richman and P. Schone, "Mining Wiki Resources for Multilingual Named Entity Recognition," in *Proc. ACL*, 2008.

[43] E. Adar, M. Skinner, and D. S. Weld, "Information arbitrage across multi-lingual wikipedia," in *Proc. WSDM2009*.

[44] G. Bouma, S. Duarte, and Z. Islam, "Cross-lingual alignment and completion of wikipedia templates," in *The Third International Workshop on Cross Lingual Information Access: Addressing the Information Need of Multilingual Societies*, 2009.

[45] G. de Melo and G. Weikum, "Untangling the cross-lingual link structure of wikipedia," in *48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), Uppsala, Sweden*, 2010.

[46] R. Navigli and S. P. Ponzetto, "Babelnet: Building a very large multilingual semantic network."

[47] W.-P. Lin, M. Snover, and H. Ji, "Unsupervised Language-Independent Name Translation Mining from Wikipedia Infoboxes," in *Proc. EMNLP2011 Workshop on Unsupervised Learning for NLP*, 2011.

[48] D. Lin, S. Zhao, B. V. Durme, and M. Pasca, "Mining parenthetical translations from the web by word alignment," in *Proc. ACL2008*, 2008.

[49] G. You, S. Hwang, Y. Song, L. Jiang, and Z. Nie, "Mining name translations from entity graph mapping," in *Proc. EMNLP2010*, 2010.

[50] L. Bentivogli, P. Clark, I. Dagan, H. Dang, and D. Giampiccolo, "The sixth pascal recognizing textual entailment challenge," in *Proc. TAC 2010 Workshop*, 2010.

[51] Z. Chen, S. Tamang, A. Lee, X. Li, W.-P. Lin, M. Snover, J. Artiles, M. Passantino, and H. Ji, "Cuny-blender tac-kbp2010 entity linking and slot filling system description," in *Proc. TAC 2010 Workshop*, 2010.

[52] V. Castelli, R. Florian, and D. jung Han, "Slot filling through statistical processing and inference rules," in *Proc. TAC 2010 Workshop*, 2010.

[53] Z. Chen and H. Ji, "A pairwise coreference model, feature impact and evaluation for event coreference resolution." in *Proc. RANLP 2009 workshop on Events in Emerging Text Types*, 2009.

[54] V. Prabhakaran, M. Bloodgood, M. Diab, B. J. Dorr, L. Levin, C. Piatko, O. Rambow, and B. V. Durme, "Statistical modality tagging from rule-based annotations and crowdsourcing," in *Proc. ACL Workshop on Extra-propositional aspects of meaning in computational linguistics (ExProM)*, 2012.

[55] R. Sauri and J. Pustejovsky, "Factbank: A corpus annotated with event factuality," in *Language Resources and Evaluation*, 2009.

[56] R. Rosales, F. Farooq, B. Krishnapuram, S. Yu, and G. Fung, "Automated identification of medical concepts and assertions in medical text," in *Proceedings of AMIA*, 2010.

[57] F. Damerau, "A technique for computer detection and correction of spelling errors," in *Communications of the ACM*, 1964.

[58] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in *Soviet Physics Doklady*, 1966.

[59] H. Ji, D. Westbrook, and R. Grishman, "Using semantic relations to refine coreference decisions," in *Proc. HLT/EMNLP 05*, 2005.

[60] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis, "Overview of the tac 2010 knowledge base population track," in *Proc. TAC 2010 Workshop*, 2010.

[61] H. Ji, R. Grishman, and H. T. Dang, "An Overview of the TAC2011 Knowledge Base Population Track," in *Proc. Text Analytics Conference (TAC2011)*, 2011.

[62] M. Richardson and P. Domingos, "Markov logic networks," in *Machine Learning*, 2006.

[63] K. Bollacker, R. Cook, and P. Tufts, "Freebase: A shared database of structured general human knowledge," in *Proc. National Conference on Artificial Intelligence (Volume 2)*, 2007.

[64] R. Grishman, D. Westbrook, and A. Meyers, "Nyu's english ace 2005 system description," in *Proc. ACE2005*, 2005.

[65] L. Bentivogli, P. Clark, I. Dagan, H. Dang, and D. Giampiccolo, "The seventh pascal recognizing textual entailment challenge," in *Proc. TAC 2011 Workshop*, 2011.

[66] A. Penas, A. Rodrigo, V. Sama, and F. Verdejo, “Testing the reasoning for question answering validation,” in *Journal of Logic and Computation*, 2007.

Assessment Models and Qualitative and Symbolic Analysis Techniques for an Electrical Circuits eTutor

Adrian Muscat
Dept of Communications and Computer Engineering
University of Malta
Msida, Malta
Email: adrian.muscat@um.edu.mt

Jason Debono
Institute for Electronics
Malta College for Science and Technology
Corradino, Malta
Email: jason.debono@mcast.edu.mt

***Abstract*—This paper is about assessment models, domain expert models and user interfaces as components in an Intelligent Tutoring System that serves junior classes in electrical circuits. Two student models for the purpose of automated assessment are developed and tested. One of the models is a Markovian graph model, while the other is a histogram model. The effectiveness of these models in tracing the student's declarative as well as procedural knowledge is studied and compared to human assessment. The domain expert models are based on qualitative analysis and on symbolic quantitative techniques. These models are used to test declarative statements made by the student and also to generate a solution to the problem. The circuit analysis techniques are also studied from an educational point of view and are compared to numerical models on the basis of how much they help the student assimilate the knowledge. Two types of user interfaces are developed, one is text command line based, and the other comprises a graphical user interface. These three building blocks are used in the development of two independent systems, which are field tested with the engagement of polytechnic teachers and students at the higher national diploma level. The technical and pedagogical results obtained for the two modules are good and encouraging.**

Keywords-Electrical; Intelligent Tutoring System; Qualitative; Symbolic; Markov Model; Assessment;

I. INTRODUCTION

Electrical circuit theory is one of the foundational courses studied in college, polytechnic and university degrees in the areas of electrical and electronics engineering. Later courses, such as electronic circuits and electrical machines, build on a good knowledge-base in circuit theory. It is therefore important that the student acquires a good handle in this theory. As with other foundational courses good mentoring from the very start is important in reaching this goal. As such Computer Aided Learning (CAL) or Intelligent Tutoring Systems (ITS) software can play a significant part in the progress of the student. In [1] the authors develop a prototype circuit simulator based on qualitative and symbolic reasoning, that emulates the process or sequence of steps that a person carrying out circuit analysis manually usually engages in. In this paper the system is augmented with the addition of an assessment module that is useful in giving feedback and following the progress of the student. These two models form the basis or kernel for an ITS or eTutor.

Personal human tutors are very effective in increasing the learning rate and studies show that personal tutoring helps students achieve significantly higher assessment scores [2] and [3]. A system of personal human tutors is however unsustainable and unrealizable due to the financial cost of the project as well as the lack of availability of human tutors. ITSs promise to deliver a personal mentor or a tutor to each student in class. The quest is to model the tutor using artificial intelligence techniques. Two early and substantially successful systems are PUMP [4] and SHERLOCK [5]. PUMP is a secondary school algebra tutor and SHERLOCK is a virtual practicing space for apprentices in electronics troubleshooting. More recent systems were designed for physics [6] and medical sciences [7], and the systems proposed and explored in [8] and [9] are probably the first ITS for electrical circuits. The system described in [8] is a production system and rules are defined to generate problems, solve problems and judge mistakes. The system generates and solves problems that consist of simple parallel and series combinations of impedances. Judging mistakes is carried out by analysing and coding several real-world student mistakes as mal-rules.

ITS were initially evaluated from an Artificial Intelligence point of view rather than from an educational impact focus. This approach is however changing and ITS is much more of an interdisciplinary research area today. Indeed today more emphasis is placed on evaluating the impact of ITS from an educational point of view. Nevertheless, the independent development of a number of components that contribute to the realisation of the ITS is a necessity.

This paper contributes a Markovian Assessment Model to assess solutions to problems in electrical circuits and a Nodal Analysis electrical circuits expert model for circuits of arbitrary topologies. Two systems that target electrical circuit classes are discussed. The first system accepts input circuits that are made up of an arbitrary number of resistors and one voltage source. This system processes serial and parallel connections of resistors to provide a machine generated full

answer and allows the student to drive the process him/herself. In the latter case, the output log-file provides the input to a student assessment model that assesses the student for both declarative and procedural knowledge. The second system accepts input circuits that are made up of an arbitrary number of resistors, voltage sources and current sources. The software tool then tutors the user on how to select valid spanning trees and the corresponding fundamental cutsets for the input circuit. The symbolic Kirchhoff's Current Law (KCL) equation for each fundamental cutset is then generated by the program, which the user or student can compare to his/her workings. Both tools analyse the topology of the input circuit to accomplish their respective type of analysis. The first system is targeted to Malta Qualifications Framework (MQF) Level Four students while the second system is to be used by MQF Level Five students.

This section introduced and motivated the need for tutoring systems in electrical engineering. The rest of the paper is organised as follows: Section II gives background information and discusses related work in the literature. In particular Section II-C reviews circuit analysis techniques and section II-D reviews student models, both of which are important in this work. The framework and models developed are described in section III. Sections IV and V describe and discuss the results. Finally, section VI concludes the paper.

II. Background and Literature Review

This section provides (a) background to the current state-of-the-art ITS architecture and the current practices in schools teaching electrical theory, and (b) a literature review of domain expert models, which in this case are circuit analysis techniques, and students assessment models that have been proposed in various ITS research projects.

A. Intelligent Tutor Architecture

Fig. 1 depicts the general architecture for an ITS, summarized from [10] and [11]. The main components of such a system are a Domain or Expert model, a Student Model and a Tutoring or Pedagogical model. The problem Solving Environment or Human Computer Interface is another component that should not be underestimated. To these components we have added a human Tutor model, which is useful to tune the ITS system to the peculiarities of specific human tutors. This involves machine learning from data generated by the human tutor and may contribute to a more effective ITS.

The problem solving environment defines the way the human student interacts with the system. It defines for example whether the interaction is via a text editor, via a graphical editor and more recently via speech and vision. The interface selected has a profound effect on the pedagogical nature of the system. It can for example limit the types of inputs that a student is allowed to enter. This can be either thought of as a limitation, in other words less freedom of roaming space for the student or as a forced scaffolded learning pedagogy. In general, the ITS research community agrees that the problem solving environment should emulate as far as possible the real

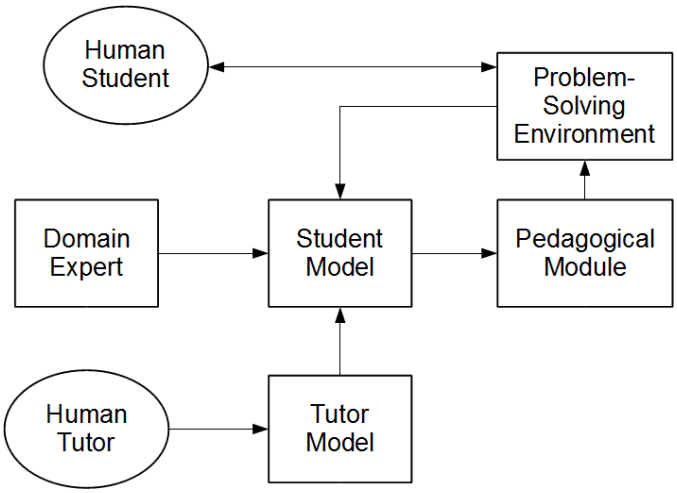


Fig. 1. Intelligent Tutoring System architecture, summarized from [10] and [11]

world environment and at the same time facilitate the learning process [10]. The latter requirement should be considered in the light that scaffolding should be completely removed at the tail end of the learning process [12].

The Domain Expert module provides an interpretation to the student's input. This module determines whether the atomic assertions of the students are correct in the specific domain area. Additionally the expert module should generate a full answer to problems given to the student, including an explanation in a natural humanistic language, symbols and diagrams included. This implies that the system must apply a causal human-like reasoning process when generating an answer. Finally, this module encompasses all the knowledge that a student is expected to learn and can therefore be termed as the *ideal student model*. In other words it is a benchmark that students strive to reach.

The student model is a record of the knowledge state of a student. In its most simple form it is a copy of the expert's model that is tagged with information of how well the student has demonstrated knowledge of each component in the expert model. Knowledge can be classified into classes; declarative and procedural [10]. Most often declarative knowledge implies learning rules and relationships. On the other hand procedural knowledge is not typically well defined and relates to the problem solving approach itself. For the case of declarative knowledge statistical models may suffice, while for procedural knowledge models that consider the sequence and order, in which declarative knowledge is applied are desired. In the electrical circuits ITS described in [9] and [8] the student model is limited to recording declarative knowledge and Mishra et al point out that the model should also capture the knowledge flow [13]. Finally, the student model is used to assess the progress of the student and its output is very useful for the pedagogical module that observes the student and controls the actions taken by the ITS.

The Pedagogical Module decides the problems and se-

quence that are presented to the student and, at which moment it offers support to the student. This model is usually considered to be domain independent. Typically there are six types of support that an ITS can provide to the students [14] and [10]; (1) Problem solving demonstration, (2) Scaffolding, (3) Monitored from a distance, (4) Goal seeking, (5) Free exploration and (6) No support is provided. The optimal choice of these six types of support services is the topic of a highly debated question in pedagogical research. In [13], Mishra et al cast this problem as an intelligent questioning system. Mishra et al argue that current electrical engineering ITS available are a "little more than a computerized version of home problems found in a typical textbook" and without an appropriate student model the pedagogical module does not function well or not at all. Another way to approach this problem of choice is to consider a domain independent help-seeking model, that aims at detecting inappropriate help requests and a gaming model that detects attempts at gaming the system. This approach is studied in depth by Roll et al in [15].

Notwithstanding the significant progress in the field of ITS, the products developed are still deemed not as effective as a human tutor in the situation of when he/she is leading discussions with students [13].

B. Electrical Theory Classes in Schools

Most college, polytechnic and university electrical circuit theory courses include theoretical as well as practical sessions. The practical sessions are important for two reasons; (a) students learn how to link theoretical models to the real-life circuits, and (b) students learn how to carry out the appropriate measurements using the right instrument. These practical skills are indispensable for professional engineers during the installation, testing and maintenance, of electrical and electronics systems. However instrumentation is generally expensive and its use is restricted to labs. In this respect numerical circuit simulators, such as SPICE, augmented with a graphical schematic capture front and back ends are very useful. With such elearning tools students connect virtual components together using virtual wires, choose and add virtual instruments to the circuit, and finally, carry out a computer analysis. The software outputs the variables chosen or measurements as displayed on the virtual instruments. Such measurements include numerical values, like for example electrical current on a virtual meter, and voltage waveforms on a virtual oscilloscope. This type of eLearning software, widely distributed among colleges and polytechnics helps students in the acquisition of practical skills including the selection of instrumentation. It also speeds up the process and reduces the cost since there is no need for building the circuit in real life. However it does not help the student in understanding how the circuit works or how to design the circuit. On the contrary, it encourages the student not to carry out a manual or mental analysis.

The second author has carried out a study based on a questionnaire regarding the effectiveness of using SPICE simulators as a pedagogical tool and using handouts, which explain step by step the symbolic calculations involved in electric circuit analysis. The questionnaire involved both open ended questions and Likert scale questions. The questionnaire was handed out to the students of the two first year classes of the National Diploma in Electrical and Electronics Engineering (MQF level 4) at Malta College of Arts, Science and Technology (MCAST), Malta and a total of thirty one filled in questionnaires were collected. The full report on this study is published in [16]. The report outlines two conclusions that are relevant to this paper; (a) in general although students find the SPICE simulator as motivating very few agree that it helps in understanding how circuits work, and (b) The larger proportion of students acknowledge that it would be much more useful if the simulator explains how the results are obtained. These results confirm what other researchers [9] and [17], who advocate the use of symbolic and qualitative techniques have stated in their papers, i.e., software that gives explanations, and not just results, is a better aid to students.

Apart from practical skills, electrical engineering students learn how to analyse and design electrical circuits. Traditionally students have been taught how to analyse electrical circuits using pen and paper through the application of the relevant theories, including Ohm's Law, Kirchhoff's Current Law and Kirchhoff's Voltage Law. As explained above SPICE simulators were not specifically designed to help students learn how circuits work, consequently SPICE simulators have some serious limitations when used as a pedagogical tool. The electrical theories taught to students are an essential part of the mental models that the students must develop. Using these theories students can write down symbolic (algebraic) equations that describe how the circuit being analysed behaves. In contrast numerical simulators calculate values iteratively, and this approach limits the understanding and insight that the simulator can impart to its user about how the circuit being analysed functions. In the last couple decades, symbolic simulators have been developed that build the symbolic equations that describe the circuit being analysed and display these equations explicitly. By examining the transfer function equations the student can then infer how the output changes when the parameter values are varied. Examples of recently developed symbolic simulators are SAPWIN [18] and SNAP [19]. On the other hand, these simulators do not explain how the transfer function is obtained.

Additionally, accomplished engineers apply mental models, during what-if analysis activities, to understand how a change in a parameter at a point of the circuit affects the other parameters of the circuit. A change in a parameter, like for example the input voltage, is thought of as influencing the parameters of its neighbouring components and nodes. In turn, these varying parameters affect their own neighbours and hence the changes propagate throughout the circuit. Furthermore, engineers only consider the direction of change, that is, an increase, a decrease or no change at all in the parameter's value. In other words, the quality of the change is considered and not the quantity of the change. This method of analysing a circuit was formally studied by Sussman and Stallman [20] and Johan De Kleer

[21]. In [9], Ahmed et.al. note that experts apply qualitative reasoning prior to calculation, whereas novices calculate first.

In summary, human tutors teach informal methods for what-if qualitative analysis and quasi-formal methods such as node and loop methods that yield quantitative answers. Likewise assessment is carried out as how well a student demonstrates the application of both qualitative analysis as well as quantitative analysis. The most creative teachers make use of real-world problems to motivate the students and include design questions to give a new meaning to circuit analyses. In [17], a Practical Relevance Module (PRM) and a Design Module (DM) are proposed for inclusion in an electrical circuit ITS.

C. Circuit Analysis Techniques

In this section, three circuit analysis paradigms, (the quasi-formal nodal and loop analysis, symbolic analysis and qualitative analysis), pertinent to ITS are discussed.

1) Ohm's Law, Nodal and Loop Analysis: The simplest way to analyse circuits is to identify parallel and series connections such that the original circuit is reduced in complexity to another equivalent circuit that is easier to analyse. However this method fails when one section of a circuit is mutually coupled to another section, in which case a global simultaneous solution is required. The electrical circuit domain model described in [8] and [9] is based on non-coupled serial and parallel combination of impedances and is there limited to such circuits. On the other hand a complex circuit can be described by using either the mesh or the nodal formulation. The mesh equations are based on Kirchhoff's Voltage Law (KVL), which states that the sum of voltage drops along any closed loop is zero, while the nodal equations are based on KCL, which states that the algebraic sum of currents leaving any node is zero. A more general definition of KCL is that in any fundamental cutset that separates the network into two parts, the sum of the currents in the cutset edges is zero. If the number of branches in the network is denoted by the letter b and number of ungrounded nodes is denoted by the letter n, then to solve a circuit; (a) the number of mesh equations required is equal to b - n, and (b) the number of nodal equations required is equal to n. In general nodal analysis yields less equations than mesh (loop) analysis and hence nodal analysis is usually easier to carry out [22].

The nodal and loop methods are widely manually applied in circuit analysis. Automation of these methods however requires formalising it in graph theory, for example as Signal Flow Graphs. The model implemented in this paper focuses on the use of graph theory to analyze the topology of electrical circuits, which is the study of inter-connected objects represented by 'edges' in a graph [22]. The points where the end-points of edges touch together are formally called 'vertices' or 'nodes'. A graph is extracted from the schematic diagram of a circuit by replacing the components with edges. For example the graph shown in fig. 2(b) is extracted from the circuit shown in fig. 2(a). A graph of an electrical circuit contains more than one spanning tree, and from each spanning tree a set of fundamental loops and fundamental cutsets can be

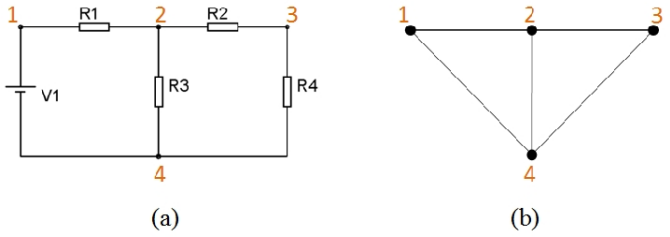


Fig. 2. (a) Example Electrical Circuit, and (b) Graph for Example Circuit.

extracted. A spanning tree of a graph is defined as any set of connecting branches that connects every node to every other node without forming any closed paths or loops [22]. Fig. 3(a)

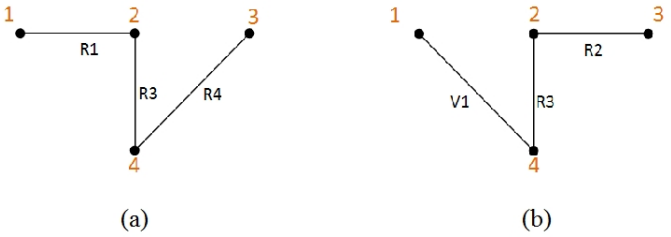


Fig. 3. (a) Spanning Tree I, and (b) Spanning Tree II.

and fig. 3(b) show two different spanning trees for the graph shown in fig. 2(b). Once a spanning tree has been defined, the edges making part of the spanning tree are referred to as branches. The remaining branches are referred to as links or chords. A fundamental loop is a loop that contains one (and only one) link in its set of edges [22]. Fig. 4(a) shows the

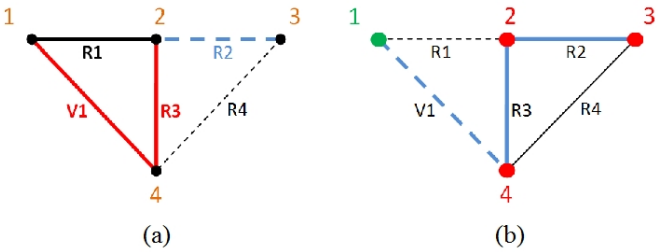


Fig. 4. (a) Fundamental Loop for R1, (b) Fundamental Cutset for V1.

fundamental loop for link R1 when considering the Spanning Tree shown in fig. 4(b). To construct a loop that includes only the link R1 (shown as a thick black line) and no other links, the tree branches shown in red must be used. Therefore the fundamental loop of R1 is made up of the edges: R1, R3, and V1. A cut set is a minimal set of edges that when cut, divides the graph into two groups of nodes. A fundamental cutset is a cutset that contains one (and only one) tree branch in its set of edges [22]. Fig. 4(b) shows the fundamental cutset for tree branch V1 when considering the Spanning Tree shown in fig. 3(b). By cutting V1 node 1, shown in green becomes isolated

from the group of remaining nodes, that is nodes 2, 3 and 4, which are shown in red. Together with branch V1, the link R1 has to be cut to keep the two groups of nodes separated, hence the complete fundamental cutset is: V1, and R1.

2) Symbolic Simulators: Symbolic simulators are based on formal circuit theory and are able to generate the transfer function of circuits input to them. The transfer function is a commonly used symbolic expression that describes how a circuit behaves. Using the transfer function the output signal that the circuit generates for a given input signal can be calculated. The advantage of using a symbolic transfer function is that the circuit is analysed symbolically only once to obtain the transfer function and then as many numerical answers as needed can be obtained from the transfer function by substituting the symbols with the numerical values being considered. Considerable research has been carried out on the symbolic analysis of electrical circuits in the late 1960's and a number of software Symbolic Simulators were developed in the 1980s [23]. For example De Kleer developed a symbolic simulator called SYN together with Sussman in 1979 [24]. De Kleer states that SYN has several limitations that are were overcome in EQUAL, the Qualitative Analysis Simulator that he developed [21]. These limitations include the lack of the ability to use approximations that drastically simplify the algebra without sacrificing accuracy. Some of these problems have been addressed in modern symbolic simulators [25]. Good examples of modern symbolic simulators that are equipped with a Graphical User Interface (GUI), including a schematic capture front end, are SAPWIN [18] and SNAP [19].

3) Qualitative Electrical Circuits Analysis: De Kleer [21] divides qualitative analysis of electrical circuits into two independent types of analysis, which are; (a) causal analysis, and (b) teleological analysis. The way that the components are connected together in a circuit gives a specific structure to the circuit. The schematic diagram of a circuit describes this structure. Each component in the circuit causes some effects on the other components that are connected to it, and these in turn affect the components that are connected to them, and so on. The aim of causal analysis is to combine the behaviour of the individual components to explain the behaviour of the overall composite system. That said, a composite system is built so that it serves a purpose. The purpose of a circuit is also referred to as the function of the circuit. Teleological analysis describes how by knowing the behaviour of a circuit one can deduce its function. Causal analysis relates structure to behaviour and teleological analysis relates behaviour to function. These two types of analysis were also investigated by Marc Fossprez in 1988 [6]. Marc Fossprez states that it is relatively easy to deduce how a circuit behaves once its function is known, but it is much harder to deduce how a circuit behaves if only the circuit's structure (its schematic diagram) is given. His work focuses on this latter task and he gives definitions about the different structures that circuits can possibly have and mathematical proofs that employ topology and graph theory.

The "Propagation of Constraints" technique, developed by Sussman and Stallman [20], is the first attempt at formalizing qualitative analysis. This method is inspired by the what-if qualitative and informal procedure applied by experts in electrical engineering and described in section II-B. The algorithm developed by Sussman and Stallman calculated the numerical values of voltages and currents and therefore goes beyond qualitative analysis. In 1984 Johan De Kleer implemented the Qualitative Analysis of electrical circuits in a program he called EQUAL [21]. This program is able to explain how a circuit works using qualitative arguments and even categorize the circuit as being a power-supply, logic-gate, amplifier or multivibrator. Furthermore, the propagation of Constraints has proved to be a powerful algorithm in circuit analysis. Fossprez recommends its use when searching for a pair of compatible current and voltage (i, v) orientations, while analysing circuits qualitatively [26]. In 2006, Rehman et al developed a type of software authoring tool for an 'Intelligent Book' [27], in which this algorithm is used to find the currents, voltages and component values inside different circuits. The values generated by the Propagation of Constraint algorithm are used to verify the values input by the students that make use of the 'Intelligent Book'.

D. Student Models For Assessment

Alongside the domain expert model, the student model is indispensable in an ITS since this stores information about the student's knowledge state, progress and learning behaviour. The two main characteristics of an ITS system is adaptivity to the student and assessment. In the case of being adaptive it is necessary to build a model for each student that is updated as the student progresses through the learning process. In the case of assessment it may not be necessary to store a model for each student. Instead models that the define certain levels of attainment, such as distinction, merit and pass, to which the student's profile is compared will suffice. In an assessment system the goal is to determine what the student knows or the knowledge state. Knowledge is often described as being either declarative, procedural, or a mixture of both. Ideally the assessment system models all three types. Additionally other variables that relate to the student's human attributes and aptitudes, such as cognitive, conative, meta-cognitive, motivational and affective can be added to the student model to deliver systems that consider hidden skills and states of mind [28] and [29]. Finally, the terms knowledge tracing and model tracing are often used to distinguish between assessment and adaptivity. Knowledge tracing refers to what the student knows, whereas model tracing refers to the steps taken by a student when solving a problem. Model tracing implies the sequence of selecting the right or wrong items versus time, that lead to a solution or an impasse. It is our view that model tracing can be split into two types, those that are directly related to the domain and therefore can be linked to the knowledge model and those that related to generic items or skills, such as self-regulation. In summary, student models are compared and contrasted on the basis of how well they can model knowledge that is either declarative, procedural, or

a mixture of both.

The pre-cursor of ITSs are Computer Adaptive Testing (CAT) systems that attempt to adjust test questions to suite the ability of the examinee. The principle behind CAT is to build and update a student model that reflects the knowledge state versus time. The choice of the next item or problem to solve is therefore based on the model output. Item Response Theory (IRT) was developed for this purpose [30],[31] and was the prevalent method of choice until the Bayesian modeling approach was extensively studied and adapted to modeling the student's knowledge state [32], [33], [34] and [35]. More recently Hidden Markov Models (HMMs) have been proposed to study the metacognitive behaviour of students during the learning process [36] and [37].

Bayesian Networks and Markovian Models are graph models that consist of nodes and directed edges. Bayesian networks are acyclic, while Markovian models allow cycles. Markovian models therefore have the property of representing knowledge in a more compact form at the expense of granularity and inference. Bayesian networks can be more complex and computationally intensive, however they have been shown to perform well in knowledge tracing and model tracing [28] whereas Hidden Markov Models have been limited to model tracing [36]. The graphical networks are characterized by two types of variables or nodes; target variables and evidence variables. Target variables are for example the knowledge state (both procedural and declarative) and cognitive skills. Evidence variables are attributes that can be measured such as answers to questions, selection of items, assertions and sequence of events including timing information. Granularity has an effect on computational time required and may have an effect on the accuracy of the judgment made on a student. Therefore, models have to be compared on this attribute. Depending on the target variables the ITS designer has to decide on what to model with a network. For example in [6], separate Bayesian networks are built for each student and for each problem presented to a student. The total number of networks that a system handles can therefore be high. Finally, in most ITS Bayesian network implementations, such as in [6] and [33] the networks are knowledge engineered and very few are constructed from empirical data. One such case is described in [35].

The work reported in [36] is motivated by an emphasis on preparation for future learning. Therefore, the goal is not to assess the knowledge state but to study strategies and behaviours that students engage in during the learning process. The proposed solution then consisted of deriving a HMM from the measured and tagged students' activity sequences. The model identifies sequence patterns to a student behaviour type or hidden state. In other words one model is required for each behavioural trait.

In this paper a Markovian Model or finite-state machine and a Histogram-based model are developed and studied in terms of how well they perform in knowledge and model tracing.

III. Implementation of Models and Framework

Two separate systems were developed. The first tool is text based and input is provided via a command-line interface. The user interface for the second tool provides a Graphical User Interface for part of the input and output, which makes it more adequate to be used as an eTutor. For both tools circuits are input via a text file. The student assessment model is integrated with the first tool only and does not provide feedback during the problem solving process. The second tool uses the expert model directly to provide feedback after each event.

A. Qualitative Analysis based Expert Model

This module is intended for entry level students and deals directly with the analysis of simple circuits made up a voltage source and a set of resistors. The software tool requires the user to input the circuit to be analysed as a text file. As such the circuit has to be specified in a matrix, in which the rows represent nodes and columns represent components. The first row of this matrix is reserved for the components' values, that is the voltage of the battery and the resistance of each resistor. In the cells of the other rows a '1' means that terminal one of the corresponding component is connected to the node corresponding to the cell's row, '2' means that terminal two of the corresponding component is connected to the node corresponding to the cell's row and '0' means that the corresponding component is not connected to the particular node considered.

The first software tool accepts input circuits that are made up of an arbitrary number of resistors and only one battery. This program then analyses the connections of all the resistors and identifies resistors that are connected in parallel. Each group of resistors connected in parallel is replaced by one equivalent resistor. The program then identifies serially connected resistors and replaces each group by one equivalent resistor. This process is depicted in fig. 5. At each step the program outputs a matrix in text format, which specifies the connections in the resultant simplified circuit. If the resultant circuit contains other groups of resistors that are connected in parallel or in series further reductions are done. This process is repeated until no further reductions are possible. The process is traced backwards and the required voltages and currents across and through the various resistors are calculated. The program is used in this mode when a demonstration or a full solution to the problem is required. In this mode, the student can then manually compare the eTutor's solution to his/her own and discover where s/he erred.

The program can be used in an interactive mode, where the student solves the problem on the eTutor console, or at least records his/her steps and calculations on the eTutor. The learner enters statements via a command-line interface, for example *Parallel(1,R1,R2,5)*, which means that in circuit 1, R1 is in parallel with R2 and the equivalent resistance is 5 ohms. The Domain Model checks that the statement is correct or incorrect. The assertion is recorded in a log-file and tagged as correct or not correct, following output from the expert model. Changes in the circuit are entered via a labeled text

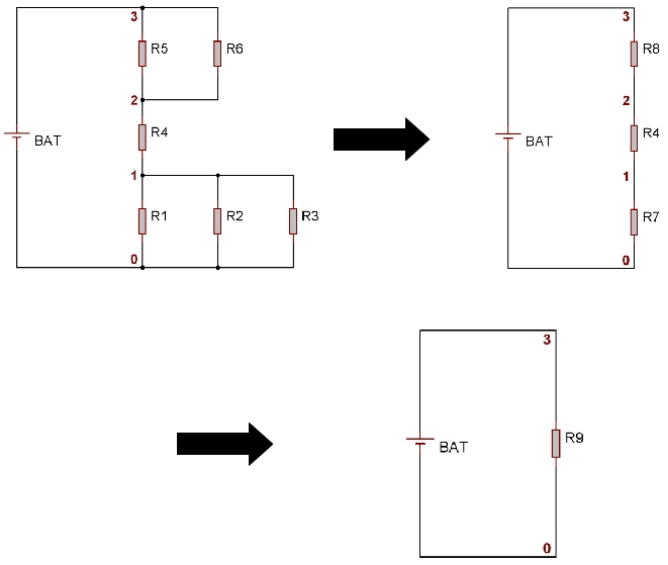


Fig. 5. Example of the parallel and series resistors reduction processes carried out by the first program.

file, whose format was described above, and the circuit is linked to the previous assertion that motivated the change in the circuit. If the same assertion is repeated, i.e. the learner back-tracks and repeats statements, the previous copies are deleted. Although deleting repeated assertions results in a loss of behavioural traces, it relaxes the task of the assessment module, which is required to output a summative value for the assessment. It also reduces the possibility of a student gaming the system. Furthermore the number of commands and inputs possible are grouped into six events. The granularity is thus reduced with some possible loss of precision. The latter step is a pre-processing step to relax the requirements from the data mining technique. It also has a positive effect of minimizing overfitting. The six events are; (a) The correct assertion of a series or parallel combination and the calculation of the equivalent resistance, (b) the incorrect assertion of the latter, (c) The correct modification of a circuit, (d) the incorrect modification of a circuit, (e) the correct assertion of a voltage or a current, and (f) the incorrect assertion of the latter. These six events labelled as (SP, $\neg SP$, C, $\neg C$, VI, $\neg VI$) are used by the models described in the next section, III-B. In summary, for the purpose of developing the assessment models described in section III-B, the sequence of events are saved to a log-file. The relationship in between events is not however preserved, although these are implicity coded in the sequence. On the other hand all the activity, including the commands and circuit scripts are saved for the purpose of manually assessing and marking the solutions. The test group for this software tool included 27 students and 3 members of the academic staff. In general the students and the academic staff think that the tool is very useful and should be expanded to include further expert knowledge. The text based user interface is a bit daunting, especially when modifying the circuit. The teachers think that the restrictions imposed by the program during the circuit analysis process can give some false expectations to students. The assessment module is described in the next section.

B. The Assessment Models

The assessment module is required to assess the performance of a student after attempting a solution to a given problem. Two different types of models are considered. One model is a Markovian Model that outputs a probability distribution over four levels of knowledge attainment and the other model is a histogram-based model that outputs a measure of difference between the ideal student model and the given case input. The six variables used in these models are the events described in the previous section, III-A and the sequence of events pertinent to a given student is obtained from the log-file described in the previous section, III-A. The following sections describe the implementation of these models.

1) The Markovian Model: Fig. 6 depicts the architecture for the model. The model consists of eight states or nodes,

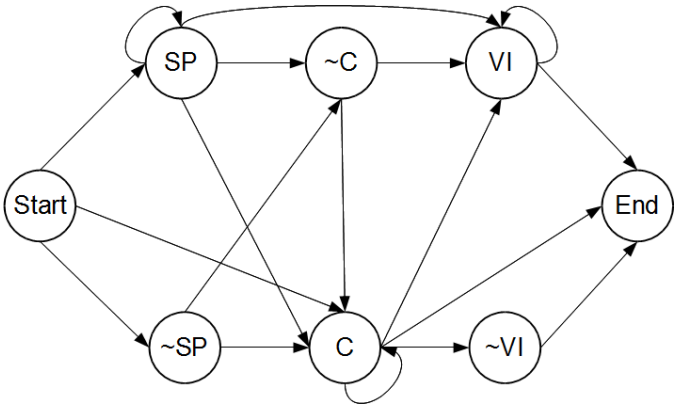


Fig. 6. The Markovian Model. Not all possible transitions are shown.

the six events described in section III-A and two new states (START and END). The value of the directed edges represent the probability of the student moving to a new state. A fully connected graph will have 64 edges in total, although the START and END states are typically only visited once. Four such models are required to assess a student in a discrete probability distribution over four levels of mastery. In this paper these four levels are labelled as 1, 2, 3, and 4, 1 being mastery of electrical circuit theory and 4 being a failure in learning electrical circuit theory. 3 and 4 are intermediate levels. The four models are all based on the same architecture in fig. 6 but the transition probabilities are different. Each case therefore models typical sequences that a student from a particular attainment level typically follows. In other words the model classifies patterns of sequential events. On the other hand atomic knowledge is assessed by the domain expert knowledge. Therefore intuitively the expert model and the Markovian Model will together achieve both knowledge and model tracing.

2) Histogram Model: The histogram-based model computes the difference between the ideal student model histogram and the case to be assessed. In the test case available there was only one ideal student model. This was due to both the nature of the problem and the fact that the human-interface module inherently restricted the freedom of the student. Two different histograms are considered. One histogram is a frequency count of the six events (SP, $\neg SP$, C, $\neg C$, VI, $\neg VI$) that a student engages in. This will be called the state vector histogram. The difference between a given case and the ideal student model is computed as,

$$\sum_i \sqrt{(S_i^{model} - S_i^{test})^2} \tag{1}$$

where $\mathbf{S}^{model}$ is the state vector histogram for the ideal student model and $\mathbf{S}^{test}$ is the state vector histogram for the given test case.

The second histogram is a frequency count of the 64 state transitions possible. This is termed the transition matrix histogram. The difference between a test case and the ideal student model is computed as,

$$\sum_{i,j} \sqrt{(T_{ij}^{model} - T_{ij}^{test})^2} \tag{2}$$

where $\mathbf{T}^{model}$ is the transition matrix histogram for the ideal student model and $\mathbf{T}^{test}$ is the transition matrix histogram for the given test case.

Since the data set includes cases that have been marked by the highest and lowest marks possible and the output for the ideal answer is zero then the results are scaled to reflect the zero to ten marks range and it will then be possible to compare the results from the histogram model to the human generated assessment. In doing this we are assuming a linear mapping.

C. Nodal Analysis Based Expert and Tutor model

The aim of this system is to eTutor students that are learning how to identify a fundamental tree and the corresponding fundamental cutsets in a given circuit and how to generate the KCL current equations for each fundamental cutset. This topic is covered in a unit called 'Further Electrical Principles' that higher national diploma (MQF level 5) students follow in the second year of their course at the MCAST.

The format of the input text file for the second program is more compact and easier to write since in it a text line is dedicated to each component and the numbers of the two nodes, to which the component is connected are stated in the corresponding line. This does away with the '0s' that were used for the first program. The other information included in each line of this text file is the X and Y coordinates of where the component is to be drawn in the GUI, the name of the component and its value.

Once a circuit is specified correctly in the input text file it can be loaded in the program. Fig 7. shows an example of a loaded circuit. The user is asked to chose a spanning tree, by clicking on the components in the circuit. Once the user selects a group of components that s/he think makes up a valid spanning tree, s/he must press the 'Check Spanning Tree' button so that the program verifies if the selected group of components makes up a valid spanning tree. If it does not the program informs the user and gives relevant feedback to the user of why the selection does not make up a valid spanning tree. The program informs the user whether s/he selected the right amount of components and whether s/he captured all the nodes in the circuit with the group of components selected. The program also informs the user if there are loops present in the selection made.

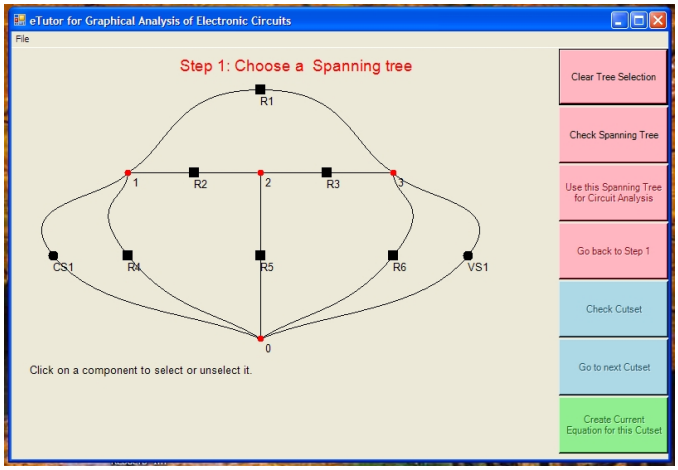


Fig. 7. Example of a loaded circuit in the program's GUI.

On the other hand, if the selection makes up a valid tree the program informs the user and allows the user to select this spanning tree to continue with the circuit analysis. To do this the user has to press the 'Use this Spanning Tree for Circuit Analysis' button. Once this button is pressed the program goes into Step 2, in which the user has to select the correct fundamental cutset for each of the tree branches inside the selected spanning tree. The tree branch, for which the user has to select the links that make up the fundamental cutset, is highlighted in red, as shown in fig. 8.

The fundamental cutset must separate one of the group of nodes from the remaining group of nodes. To help the user the program highlights all the nodes in one of these groups in orange and the nodes in the other group in green. After that the user selects the components that s/he thinks make up the fundamental cutset, s/he must press the 'Check Fundamental Cutset' button. Once this button is pressed the program checks if the selected components make up a valid fundamental cutset. If this is not the case the program gives relevant feedback to the user. The program states whether one or more components that should be included in the selection are not selected and it also states if one or more components that should not be included in the selection are in fact selected. In the case when the selected components make up a valid fundamental cutset, then the user is informed accordingly and is allowed to press the button labelled "Create Current Equation for the Cutset". When this button is pressed the KCL equation for the

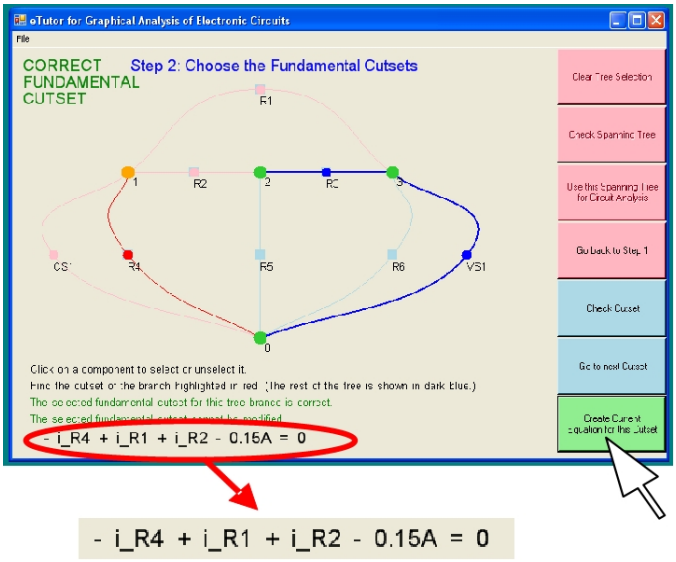


Fig. 8. Example of a fundamental cutset KCL equation generated when the correct fundamental cutset is selected and the appropriate button is pressed.

fundamental cutset is generated by the program and displayed at the bottom of the screen as shown in fig. 8. The user can then press the 'Go to next Cutset' button to find the fundamental Cutset of the next branch in the spanning tree. This process has to be repeated until the fundamental cutsets of all the branches in the spanning tree are found. At this point it is desirable that the program tutors the user on how to find the fundamental loop for each link present in the graph, but this feature has not been implemented yet. The author plans to have this feature functional in the future so that it can be used by the higher national diploma students.

In the system developed, the nodes are implemented in a list data structure. The branches or components at a given node are defined in another list. The algorithms then operate on these lists. From graph theory it is known that a valid spanning tree must be made up of n-1 edges, where n is the number of nodes. Hence the first check made to verify the input tentative spanning tree is to count the number of selected components and check if it equal to n-1. If this is not the case it means that the selected components do not make up a valid spanning tree.

The next step to carry out is to check that the selected components capture all the nodes inside the circuit (graph). The algorithm just has to go through all the selected components and mark the two nodes, to which each component is connected as captured. After that the algorithm has to go through the nodes and check that none of them is non-captured. If one or more nodes are non-captured then the selected components do not make up a valid spanning tree. There exist cases, in which the two checks explained above are satisfied but the selected components still do not make up a valid spanning tree. In this case the selected branches will not be continuously connected and at least one loop will be present in the selection. To check for such cases the spanning tree algorithm starts off with one of the selected tree branches. It checks, to which nodes this branch is connected and proceeds to discover the other branches that one of these nodes is connected to. If there are more than one branch connected to this node the algorithm starts considering the first branch and it takes note, of which branch this is so that once it finishes checking it and returns to the last node considered, it continues looking for the correct branch. This process is repeated for each node. When at least one branch is found connected to a node the algorithm jumps to the other node, to which this branch is connected and hence travels further away from the first node that it considered at the start. Naturally the larger the selected tentative spanning tree is, the more searching the algorithm has to do. But in the case of invalid spanning tree selections there are two possible ways, in which the algorithm completes. One way is that the algorithm steps forward (not backwards) into a node that it already checked, and hence a loop is discovered. The other way, in which the algorithm can complete in the case of an invalid spanning tree selection, is that it finds out that it exhausted all the branches and nodes that are connected to the first branch considered, but it did not find all the nodes present inside the graph. In this case it means that the algorithm has found one continuous length of connected branches, which is not connected to the remaining branches of the selected tentative spanning trees. Since spanning trees should not contain any discontinuities in their branches' connection, this means that the selected components do not make up a valid spanning tree.

Another algorithm used in the graphical analysis program is the one that highlights in different colours the two groups of nodes that are to be separated by a fundamental cutset. The searching that this algorithm does is very similar to that done by the algorithm that verifies spanning trees. However in this case, the fundamental cutsets algorithm does not check for loops because it is used after that a valid spanning tree is already selected, so it is already guaranteed that no loops are present. The important feature that this algorithm possesses, similarly to the previous algorithm, is that it always remembers, which branch it checked last when jumping from one node to another, so that when it returns back to the node from where it jumped, it continues checking from the correct branch.

IV. Results

This section describes results pertaining to (a) the Markovian Assessment Model, (b) the Histogram Assessment Model, and (c) the Problem Solving Environment.

A. The Markovian Assessment Model

In this paper, the Markovian models are tuned using empirical data. For this purpose twenty-seven students are given a problem to solve and their solution is recorded as described in section III-A. A human tutor then assesses the twenty-seven solutions and marks them over a 21 point scale from 0 to 10. The data set is clustered on these marks as; marks equal or above 7.5 corresponding to level 1, marks in between 5.0

and 7.4 as level 2, 2.5 to 4.9 as level 3 and marks less than 2.5 as level 4. The twenty-seven cases where approximately uniformly distributed across the four levels. Finally these groups are used to find the transition probabilities for each of the four models. Assessment consists of first computing the log likelihood probability distribution for a given new student and the case is classified by choosing the maximum log likelihood. The Markovian Model was first tested using samples from the data set. Fig. 9 shows the classification results for all twenty-seven cases, out of which two were incorrectly classified. Fig. 10 depicts the probability distributions for four correct cases and the two incorrectly classified cases. To further test the model, the twenty-seven test data set was split into two data sets, one being the tuning set and the other being the test set. The ratio of the tuning set size to the test set size was varied. When the training set size is 27 the correct classification rate is 93%, for a training set size of 23 the correct classification rate is 89%, for a training set size of 19 the correct classification rate is 85%, for a training set size of 15 the correct classification rate of 85% and for a training set size of 8 the correct classification rate is 70%.

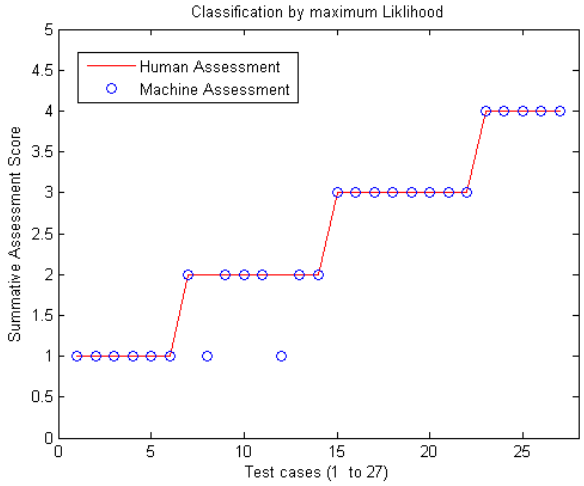


Fig. 9. Classification of results from the Markovian model by maximum likelihood.

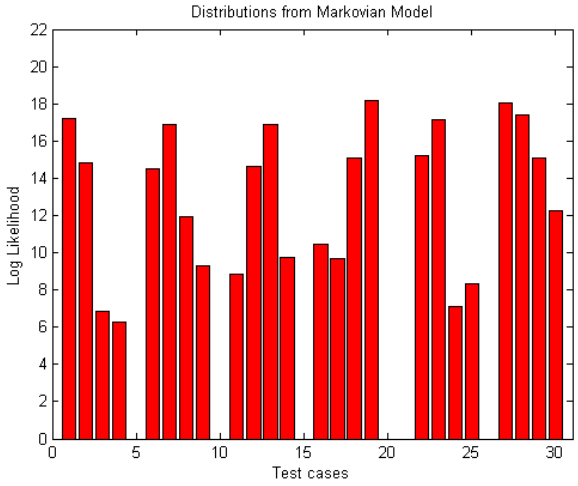


Fig. 10. Probability distributions obtained from Markovian model. The first four groups are correctly classified cases for each assessment score level. The last two groups are the two out of twenty-seven incorrectly classified cases.

is accurate only for cases close to the ideal student model, fig. 12.

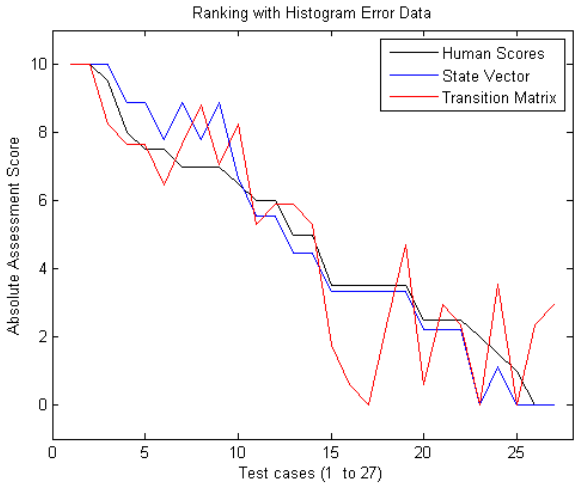


Fig. 11. Absolute assessment scores obtained from the state vector histogram and transition matrix histogram.

B. The Histogram Assessment Model

Fig. 11 shows how the two histogram models compare with the human generated assessment. A non-monotonically decreasing graph indicates deviations from the human assessor. The state vector model is characterized by a closer match than the transition matrix histogram model. This is expected since the former tests for the correct event selection and its frequency of selection, whereas the latter tests for the transitions. These results omit the states that describe an incorrectly executed event. This makes a fair comparison since we know that the human assessor did not negatively mark the solutions. When these events are included in the histogram model the deviations and oscillations increase and the model

The real number outputs from the histogram models are used to classify the students into one of the four levels of attainment. Fig. 13 shows that the classification results are not very good. The state vector model classified eight instances in the wrong class, while the transition matrix model classified twelve instances in the wrong class. Fig. 14 shows the classification results for the model that included negative states. In the case of the state vector model ten instances are not correctly, whereas for the transition matrix model eleven instances are in error. In summary, the histogram models can only be used to classify students in two states, mastery or non-mastery. On the other hand, the Markovian Model grades the

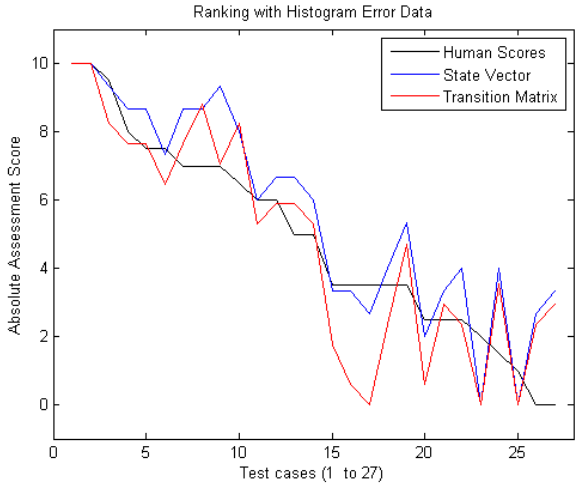


Fig. 12. Absolute assessment scores obtained from the state vector histogram and transition matrix histogram. These results include the negative states.

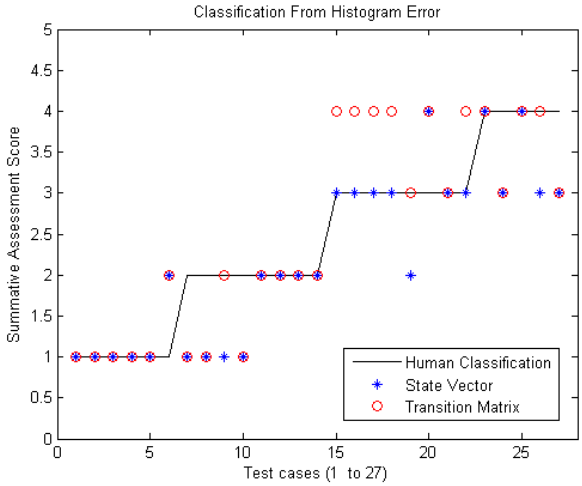


Fig. 14. Linear classification of test data using the state vector histogram and transition matrix histogram. These results include the negative states.

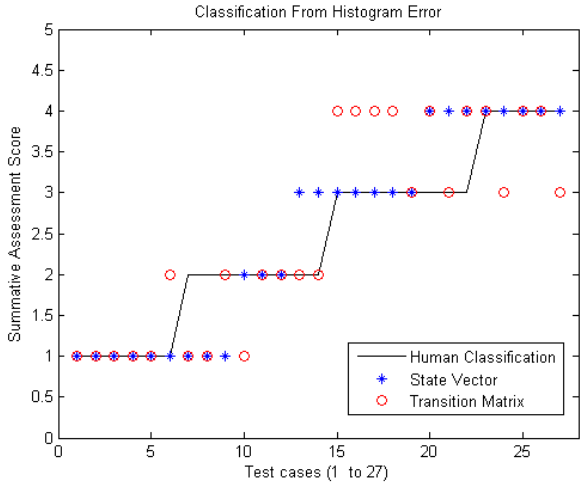


Fig. 13. Linear classification of test data using the state vector histogram and transition matrix histogram.

student on a scale of four points.

C. Problem Solving Environment

The Nodal Analysis Based Tutor System is aimed at second year higher national diploma (MQF level 5) students. Its aim is to tutor these students on how to find correct spanning trees and fundamental cutsets in graphs of electrical circuits. This program was tested and verified to function correctly. It interacts with its users through a GUI. It lets the user input the circuit of interest and select a valid spanning tree. When a valid spanning tree is selected the program lets the user work out all the valid fundamental cutsets corresponding to this spanning tree and then generates the corresponding KCL equations. Whenever the user does an incorrect choice during the selection process, the tool explains why the choice is incorrect, and hence acts like a Tutor. The tool was first tested by fifteen students that undertook courses that included the topic under consideration in the previous academic year and all these students stated that this tool would have been of great help to them. The program was then tested with a class of 18 novel HND students. These students were able to choose their own personal set of branches that make up valid spanning trees and fundamental cutsets in class. This reduced the amount of time that the students needed to learn and understand these two concepts, as well as the success rate among students.

V. Discussion of Results

The field studies demonstrated that students are keen to experience software tools that help them learn how to analyze electrical circuits in the same way a human tutor would do. Two systems that give explanations of the analysis carried out on circuits were developed. The expert models in both systems are based on symbolic and qualitative analysis, and feedback is provided on declarative and procedural knowledge. These systems therefore simulate the full-time availability of a tutor and immediate assessment results given to students increased their motivation to discover and learn. These characteristics can have a significant impact on attainment levels for a large number of students. Besides being used by one of the authors, this program was demonstrated to two lecturers that teach circuit theory and both are of the opinion that this program will help them deliver the concerned topic more efficiently, leading to higher success rates among students.

The text-base system, whose expert model is based on Ohm's Law, is targeted towards first year national diploma (MQF level 4) students. The expert model identifies resistors that are connected in parallel and in series and replaces them by equivalent resistors. Alternatively the process of analysis is carried out by the student and the system responds with immediate feedback on every assertion. Students reported, that the expert explanation provided by the system is very useful. However, the text-based interactive environment is not

straightforward and the students had to adapt to it. The main problem with this environment relates to interpretation errors specifically when the student needs to translate text into a circuit diagram, either mentally or on a side paper note-pad. Augmenting the system with a graphical view of the circuit in question would therefore mitigate this problem. Other than the latter shortcoming, both the students and teachers liked the tool and think that it is a useful tutor.

The user interface for the Nodal Analysis based expert model is based on a graphical layout. So it is not surprising that students and teachers found it easier to use. The students and teachers noted that sometimes the system either gives too much feedback or the feedback is too wide and not selective. This is understandable since no personalized student model was integrated in this system. It also shows, how important it is for an ITS to keep the student motivated and interested. Nonetheless field tests showed a reduction in the amount of time that the students needed to learn and understand the concepts in electrical circuit theory classes, as well as an increase in the success rate among students, when subjectively compared to previous groups that did not use the system.

An important aspect of the contribution in this paper is the Markovian assessment model that traced both declarative and procedural knowledge components in the student solution. The main drawback of the Markovian model is the fact that a training set is required for every problem set by the human tutor. It may be possible to generate the ideal student model from the solution generated by the domain model. Perturbations from the ideal model can then generate inferior answers to the problem and the models are tuned or fitted with simulated data. It may also be possible to use past answers and sample human generated assessments to generate synthetic answers and assessments. This solves the problem of requiring the human tutor to correct a sample class to tune the model with.

The student assessment models reported in this paper were deployed to provide feedback any time the student engages in a learning activity. This means that these systems can be used to assess students more often, providing valuable data to help teachers allocate resources more effectively and also helps in tackling challenges in mixed-ability classes. Additionally the Markovian Model is trained using human assessment data and this means that the model can simulate specific characteristics of teachers. This leads to a personalised teacher machine assistant, which learns how to assess from the human teacher.

From this experience, we see three areas that are worth improving. It would be ideal to have machine tunable Markovian assessment models. The two domain expert models should be merged into one, yielding a model that can cover most of the electrical circuit theory dealing with linear components. A student model for each and every student, possibly a probabilistic one, should be included in order for the system to provide personalized and more appropriate feedback.

VI. Conclusions

A Markovian Assessment Model and a Nodal Analysis electrical circuits expert model for circuits of arbitrary topologies have been developed and tested using lab and field tests. These two contributions are a significant improvement over the respective models described in [8] and [9].

The Markovian Model traces declarative and procedural knowledge in solutions to problems in electrical circuits. A simpler histogram model that traces only declarative knowledge is developed and compared to the Markovian Model. The Histogram Model is useful to test whether the student has mastered the topic in question and is similar to models installed in current electrical circuits ITSs [13]. The Markovian model can be possibly improved, in terms of providing finer granularity in assessment, by feeding a regressor with the four element vector that is ouput from the Markov Model.

The circuit expert model based on the formal theory of nodal analysis and on qualitative assertions is suitable for analyzing a circuit of arbitrary topology and for providing a detailed account of the analysis process. The nodal analysis based model is an improvement over the simpler symbolic and qualitative circuit model based on Ohm's law implemented in [8] and [9]. From an electrical circuit theory ITS system point of view it would be ideal to combine the features of both models into one. Other ITS or CAL systems, described in [27] and [17], make use of "the propagation of constraints" algorithm to calculate circuit parameter values. However this method does not yield an explanation as one would expect from a human tutor. So, alternatively further research may develop the "propagation of constraints" model to be better suitable for an ITS.

Finally, field tests confirmed two important points. During learning, domain models based on qualitative and symbolic analysis are more effective than simulators based on numerical analysis, which may be better suited for expert use in industry. Students prefer a problem solving environment that comprises both text and graphics-based input/output systems.

References

[1] J. Debono and A. Muscat, "An electrical circuits e-tutor based on symbolic and qualitative analysis," in *The Fifth International Conference on Advanced Engineering Computing and Applications in Sciences, ADVCOMP2011*, November 2011.

[2] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to- one tutoring.," *Educational Researcher*, vol. 13, pp. 4–16, 1984.

[3] P. A. Cohen, J. A. Kulik, and C. C. Kulik, "Educational outcomes of tutoring: A metaanalysis of findings.," *American Educational Research Journal*, vol. 19, pp. 237–248, 1982.

[4] J. R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive tutors: Lessons learned," *The Journal of the Learning Sciences*, vol. 4, pp. 167–207, 1995.

[5] A. Lesgold, G. Eggan, S. Katz, and G. Rao, "Possibilities for assessment using computer-based apprenticeship environments," in *Cognitive Approaches to Automated Instruction* (J. Regian and V. Shute, eds.), (Hilisdale, NJ), Lawrence Eribaum Associates, 1992.

[6] C. Conati, A. S. Gertner, and K. VanLehn, "Using bayesian networks to manage uncertainty in student modeling," *User Modeling and User-Adapted Interaction*, vol. 12, no. 4, pp. 371–417, 2002.

[7] S. Suebnukarn and P. Haddawy, "Modeling individual and collaborative problem-solving in medical problem-based learning," *User Modeling and User-Adapted Interaction*, vol. 16, no. 3-4, pp. 211–248, 2002.

[8] A. Yoshikawa, M. Shintani, and Y. Ohba, "Intelligent tutoring system for electric circuit exercising," *Education, IEEE Transactions on*, vol. 35, no. 3, pp. 222–225, 1992.

[9] M. Ahmed and M. Bayoumi, "An artificial intelligent instructor for electrical circuits," in *Circuits and Systems, 1994., Proceedings of the 37th Midwest Symposium on*, vol. 2, pp. 1362 –1365 vol.2, aug 1994.

[10] A. T. Corbett, K. R. Koedinger, and J. R. Anderson, "Intelligent tutoring systems," in *Handbook of Human-Computer Interaction* (M. G. Helander, T. K. Landauer, and P. Prabhu, eds.), vol. 37 of *2*, (Amsterdam, The Netherlands), Elsevier Science, 1997.

[11] H. S. Hwana, "Intelligent tutoring systems: An overiew," *Artificial Intelligence Review*, vol. 4, pp. 251–277, 1990.

[12] R. R. V. D. Stuyf, "Scaffolding as a teaching strategy," *Adolescent Learning and Development*, November 17 2002.

[13] M. Mishra, V. Mishra, and H. Sharma, "Intellectual ability planning for intelligent tutoring system in computer science engineering education," in *Emerging Trends and Applications in Computer Science (NCETACS), 2012 3rd National Conference on*, pp. 26–30, IEEE, 2012.

[14] D. M. Towne and A. Munro, "Supporting di verse instructional strategies in a simulationoriented training environment," in *Cognitive Approaches to Automated Instruction* (J. Regian and V. Shute, eds.), (Hilisdale, NJ), Lawrence Eribaum Associates, 1992.

[15] I. Roll, R. Baker, V. Aleven, B. McLaren, and K. Koedinger, "Modeling students' metacognitive errors in two intelligent tutoring systems," *user modeling 2005*, pp. 151–151, 2005.

[16] J. Debono, "Effectiveness of using circuit analysis software in vocational electronics engineering courses," tech. rep., Malta College of Arts, Science and Technology (MCAST), Corradino, Malta, September 2010.

[17] R. Amarin, K. Sundaram, A. Weeks, and I. Batarseh, "Importance of practical relevance and design modules in electrical circuits education," in *Global Engineering Education Conference (EDUCON), 2011 IEEE*, pp. 792–796, IEEE, 2011.

[18] A. Luchetta, S. Manetti, and A. Reatti, "Sapwin - a symbolic simulator as a support in electrical engineering education," *IEEE Transactions on Education*, vol. 44, p. 9, May 2001.

[19] D. Biolek, "Snap - program with symbolic core for educational purposes," *Proceedings of 4th World Multi-Conference on: Circuits, Systems, Communications and Computers*, pp. 1711–1714, July 2000.

[20] G. Sussman and R. Stallman, "Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis," *Artificial Intelligence*, vol. 9, pp. 135–196, October 1977.

[21] J. de Kleer, "How circuits work," *Artificial Intelligence - Special volume on qualitative reasoning about physical systems*, vol. 24, pp. 205–280, December 1984.

[22] J. W. Nilsson and S. A. Riedel, *Electric Circuits*. Addison Wesley, 5th ed., 1996.

[23] G. Gielen, P. Wambacq, and W. Sansen, "Symbolic analysis methods and applications for analog circuits: A tutorial overview," *Proceedings of the IEEE*, vol. 82, pp. 287–304, 1994.

[24] J. de Kleer and G. Sussman, "Propagation of constraints applied to circuit synthesis," *International Journal of Circuit Theory and Applications*, vol. 8, pp. 127–144, 1980.

[25] H. Floberg, *Symbolic Analysis in Analog Integrated Circuit Design*. Kluwer Academic Publishers, 1997.

[26] M. Fossprez, *Qualitative Analysis of Non-linear, Non-reciprocal Circuits*. John Wiley and Sons, 1992.

[27] K. Rehman, W. Billingsley, and P. Robinson, "Writing questions for an intelligent book using external ai," *Proceedings of the Sixth IEEE International Conference on Advanced Learning Technologies*, pp. 1089 – 1091, 2006.

[28] E. Milln, T. Loboda, and J. L. P. de-la Cruz, "Bayesian networks for student model engineering," *Computers and Education*, vol. 55, no. 4, pp. 1663 – 1683, 2010.

[29] R. S. Baker, "Mining data for student models," in *Advances in Intelligent Tutoring Systems* (R. Nkmabou, R. Mizoguchi, and J. Bourdeau, eds.), (Secaucus, NJ), pp. 323–338, Springer, 2010.

[30] A. Birnbaum, "Some latent trait models and their use in infering an examinee's ability," in *Statistical theories of mental test scores* (F. Lord and M.Novick, eds.), pp. 397–472, Addison-Wesley, 1968.

[31] W. J. van der Linden, *Hanbook of modernitem response theory*. Springer-Verlag, 1997.

[32] M. C. Desmarais, A. Maluf, and J. Liu, "User-expertise modeling with empirically derived probabilistic implication networks," *User Modelling and user-Adapted Interaction*, vol. 5, no. 3-4, pp. 283–315, 1995.

[33] J. Martin and K. Vanlehn, "Student assessment using bayesian nets," *Int. J. Human-Computer Studies*, vol. 42, pp. 575–591, 1995.

[34] E. Millan, M. Trella, J.-L. P. de-la Cruz, and R.Conejo, "Using bayesian networks in computerized adaptive tests," in *Computers and Education in the 21st Century* (M. Ortega and J. Bravo, eds.), pp. 217–228, Kluwer, 2000.

[35] J. Vomlel, "Bayesian networks in educational testing," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 12, no. Supplement-1, pp. 83–100, 2004.

[36] H. Jeong and G. Biswas, "Mining student behavior models in learning-by-teaching environments," *Educational Data Mining*, 2008.

[37] H. Jeong, G. Biswas, J. Johnson, and L. Howard, "Analysis of productive learning behaviors in a structured inquiry cycle using hidden markov models," *Educational Data Mining*, 2010.

Automated IT Management using Ontologies

Andreas Textor, Fabian Meyer, Reinhold Kroeger
Distributed Systems Lab
RheinMain University of Applied Sciences
Unter den Eichen 5, D-65195 Wiesbaden, Germany
{*firstname.lastname*}*@hs-rm.de*

***Abstract*—For the management of IT systems, numerous models, protocols and tools have been developed. To achieve the long-term goal of comprehensive, highly automated IT management, the various sources of information need to be combined. As syntactic translation is often not sufficient, ontologies can be used to unambiguously and comprehensively model IT environments including management rules. In this paper, we present an approach that combines the domain model, rules, instance data (which represents real-world systems) into an ontology. As the basis for an IT management ontology, we convert the Common Information Model (CIM), a Distributed Management Task Force (DMTF) standard, into an OWL (Web Ontology Language) ontology. Moreover, probabilistic knowledge of the domain is modeled using Bayesian networks and integrated into the ontology. Furthermore, the approach describes a runtime system that merges monitoring data into the ontology and then uses a reasoner to evaluate management rules.**

***Keywords*-ontology; IT management; CIM; Bayesian network**

I. Introduction

In the domain of IT management, numerous models, protocols and tools have been developed. Notable models include the OSI (Open Systems Interconnection) network management model (also known as Common Management Information Protocol (CMIP)) and the still widely used simple network management protocol (SNMP). A more recent approach to specify a comprehensive IT management model is the Common Information Model (CIM, [2]), a widely recognized Distributed Management Task Force (DMTF) standard. The more complex an IT environment gets, the more important the capability becomes to automate as many tasks as possible. Both commercial and free management tools and frameworks exist that cover different parts of the required feature set for management tasks, but usually not only a single tool, but a set of tools is used. In order to achieve a unified view of the heterogenous integrated management models, mappings between different types of models can be defined. However, syntactic translations are often not sufficient, when the same concept is represented differently in multiple domains. This problem can be approached by using ontologies to clearly define the semantics.

Only when a comprehensive formal representation of the domain data exists, that is also capable of modeling rules, a largely automatic management becomes possible, because then not only structural, but also behavioural information can be expressed in the model. To achieve such an automated management system, we describe a runtime system that imports the corresponding domain model into the ontology and evaluates the rules, based on up to date monitoring data from the system under management. In order to represent the monitoring data in the ontology, instance data is acquired at runtime and added to the ontology, so that rules can be evaluated by a reasoner according to both model and instance data.

The approach presented in this paper uses an OWL (Web Ontology Language, [3]) ontology to combine the domain model, instance data and rules defined in SWRL (Semantic Web Rule Language) in order to create a system that can automatically manage an IT environment. This results in a comprehensive knowledge base that includes both the statically loaded domain model and dynamically updated runtime information about the system under management, as well as rules to control the behavior of the system. To model entities and relationships of an IT environment, the CIM model was converted into an OWL ontology.

A domain as complex as IT management cannot be modeled solely using exact and complete information, which might not be available. Instead, probabilistic modeling and evaluation might be adequate. To enable that, the ontology and the runtime system need to be extended accordingly. As neither CIM nor an OWL ontology have native facilities for the representation of such information, Bayesian Networks are employed. Bayesian networks are probabilistic models to specify causal dependencies between random variables in a directed acyclic graph. To model probabilistic knowledge, ontology elements are annotated so that a Bayesian network can be partially derived from the ontology at runtime.

Section II gives a short introduction on the Common Information Model and describes related work in the context of ontologies and IT management. The concepts for the translation of CIM into an OWL ontology and the concepts for the combination of Bayesian networks with an ontology are described in Section III. Section IV gives an overview of our architecture for the runtime system, that is based on the aforementioned concepts for automated IT management. The paper draws a conclusion in Section V.

II. Related Work

A. The Common Information Model (CIM)

This section briefly describes the basic properties of the Common Information Model [2]. CIM is an object-oriented model that describes the entities in an IT environment and the relationships between them. This covers both hardware and software entities. The goal is to comprehensively model every aspect that is needed for consistently monitoring and managing the IT environment. CIM consists of three parts:

- A basic information model called the *meta schema*. The meta schema is defined using Unified Modeling Language (UML, [4]).
- A syntax for the description of management objects called the *Managed Object Format* (MOF).
- Two layers of generic management object classes called *Core Model* and *Common Model*.

Figure 1 shows the CIM meta schema definition in UML, from the CIM specification [2]. The meta schema specifies most of the elements that are common in object-oriented modeling, namely

- *Classes*, *Properties* and *Methods*. The class hierarchy supports single inheritance (generalization) and overloading of methods. For methods, the CIM schema specifies only the prototypes of methods, not the implementation.
- *References* are a special kind of property that point to classes.
- *Qualifiers* are used to set additional characteristics of Named Elements, e.g., possible access rules for properties (`READ`, `WRITE`), marking a property as a key for instances (using `Key`) or marking a class as one that can not be instantiated. Qualifiers can be compared to Java annotations; some qualifiers also have parameters.
- *Associations* are classes that are used to describe a relation between two classes. They usually contain two references.
- *Triggers* represent a state change (such as create, delete, update, or access) of a class instance, and update or access of a property.
- *Indications* are objects created as a result of a trigger. Instances of this class represent concrete events.
- *Schemas* group elements for administrative purposes (e.g., naming).

Properties, references, parameters and methods (method return values) have a data type. Datatypes that are supported by CIM include {s,u}int{8,16,32,64} (e.g., uint8 or sint32), real{32,64}, string, boolean, datetime, and strongly typed references (<classname> ref).

In addition to the CIM schema, CIM specifies a protocol, based on XML over HTTP, which is used by CIM-capable managers to query classes, instances and invoke methods against a so-called CIM object manager (CIMOM).

B. Ontologies in IT Management

There are several publications that examine the application of ontologies to the domain of IT management, e.g., [5], [6], [7]. The general consensus is that OWL is well suited for the modeling of IT systems, as it provides powerful modeling capabilities paired with the ability to formulate rules as part of the model and the ability to modularize the ontology. Still, both the complete translation or mapping of existing IT management models into OWL and the creation of an ontology-based automated IT management system are no solved problems.

In [5] the authors provide mappings for parts of different IT management models to OWL, including Structure of Management Information (SMI) and the Common Information Model (CIM). The resulting ontology can be used to combine the knowledge given in the different representations into a joint model. One problem the authors point out for the mapping is information that can be expressed in the original languages, but has no direct representation in OWL, such as the attachment of measurement units or access authorizations to properties. To solve this problem, the data is presented on the Resource Description Framework (RDF) layer of OWL. In RDF, it is possible to attach additional information to edges in the graph so that the data can be represented. However, this information is not available for evaluation by an OWL reasoner and therefore this approach has only limited use for automation that relies on the evaluation of rules from the ontology.

[6] describes how to represent several abstraction layers of a system in split ontologies to achieve a pyramid-like structure of ontologies, where often used ontologies are at the bottom of the figure. The reuse of components and models is an important topic in IT systems, and especially for ontology-based automation. The paper shows that OWL is capable of organizing several abstractions of a system in ontologies and reuse defined components in higher layers. This is an important aspect for the realization of a real-world management system.

A real-world management application is shown in [7] where ontologies are used to manage a network infrastructure. SWRL rules are used to create new object property connections between entities in case of a blackout. For this, properties and instance structures are observed. As a basis, Policy-based Network Management (PBNM) [8] was used. Rules are evaluated periodically during runtime, and new facts are added to the ontology. A management component observes the ontology and maps newly added facts to management operations to adjust the system.

In order to create a comprehensive ontology to model the system under management, a suitable domain model is required. The Common Information Model was examined in several publications (e.g., [9], [10]) and is often proposed as a domain model for the IT management domain, but the

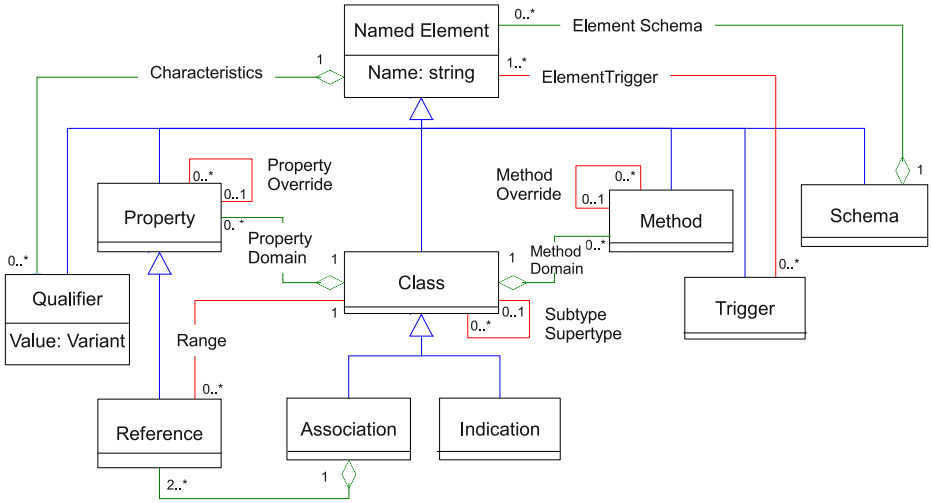


Figure 1. CIM Meta Schema [2]

authors in [9] show that it is a semi-formal ontology that has limited abilities for knowledge interoperability, knowledge aggregation and reasoning. In practice, this means that it is difficult to combine it with models from other domains, that it has no features for reasoning or the definition of rules as part of the model, and that it has only very limited built-in querying capabilities.

One solution to overcome the shortcomings of CIM while still benefiting from the comprehensive model is to translate CIM into a standard ontology format. In [9] the authors compare possible conversions of CIM to RDFS (Resource Description Framework Schema) and to OWL. They find that RDFS is unsuitable to express CIM as it does not allow to express constructs such as cardinality restrictions and some CIM qualifiers. In [5] the authors provide a possible mapping for a subset of CIM to OWL and the authors in [10] introduce a meta-ontology to model CIM constructs that have no direct OWL correspondence, but they do not describe how this meta-ontology is constructed, and their approach does not specify how several qualifiers and more complex elements, such as CIM methods, can be converted.

We developed a full mapping of CIM to OWL, which we use in this paper (see Section III-A) and which is first described in [11].

C. Bayesian Networks in IT Management

Bayesian Networks are used for the prediction of states for unobservable random variables. In IT systems management the probabilistic model of Bayesian Networkds suits root cause analysis and failure prediction well.

In [12] Bayesian Networks are used for hard disk failure prediction. Recorded Self-Monitoring, Analysis and Reporting Technology (SMART) data is used to generate the conditional probabilities for the nodes of the network. Two methods are used for the learning process, a clustering methods for sub model extraction and Expectation Maximization (see [13]) and Supervised naive Bayes Learning (see [14]). After the learning process, the SMART characteristics of a hard disk can be set as evidence in the network and a prediction for the probability of an upcoming failure can be made.

[15] considers the reliability for software systems. A special form of Bayesian Networks, the Markov Bayesian Networks, is used to predict failures in discrete time systems. Especially the working profiles of the system defined in [16] are considered as input parameters for the network. The conditional probabilities are extracted from software metrics. It is shown that the model does not provide optimal predications for failure rates, but better predictions than the Discrete Time Hyperexponential Model for Software Reliability [17].

[18] shows another attempt to predict software failures with Bayesian Networks. Complex internal relationships and correlations are considered to identify factors for failure contributions of components. For the construction of the network, a directed graph is generated where all variables are connected. Only boolean state spaces are allowed for variables. Every edge is weighted with its probability and a maximum weighted spanning tree is generated. By the choice of a root node a directed graph is achieved. Test data of the Eclipse Project [19] was used to show that the model can make statistically significant assertions about the failure probability.

D. Combination of Ontologies and Bayesian Networks in IT Management

There are no methods known to the authors for the combination of ontologies and Bayesian Networks in an IT Management context, but there are approaches to embed probabilities into OWL. In [20] the embedding of proba-

bilistic knowledge for OWL class membership is presented. The major problems are the representation of probabilistic knowledge in OWL, the derivation of an acyclic graph and the construction of the conditional probabilities. Therefore, special OWL classes are defined to represent the expressions $P(A)$, $P(A|B)$ and $P(A|\overline{B})$, which have properties for conditions, values and probabilities. These properties are used to generate the conditional probabilities. A specially modified reasoner is needed to evaluate the ontology, so that existing reasoners cannot be used.

[21] defines the language PR-OWL. In contrast to [20], no existing language constructs of OWL are used, but a proprietary extension is defined. A probabilistic ontology must define at least one individual of the special class "MTheory". A theory consists of a set of fragments that can be either context, residuum or input. Every random variable is a node with a set of possible values and conditional probabilities. With a special reasoner these constructs can be evaluated.

E. Belief Change

The process of changing beliefs to take into account a new piece of information about the world is called *belief change*. Belief change studies the process an agent has to perform to accommodate new or more reliable information that is possibly inconsistent with existing beliefs. Usually, beliefs are represented as a set of sentences of a logical language.

In the literature, three types of belief change operations can be distinguished: contraction, expansion and revision. Contraction is retracting a sentence from a belief set, expansion is adding a sentence to a belief set regardless of whether the resulting belief set is consistent or not, and revision is incorporating a sentence into a belief set while maintaining consistency. Alchourrón, Gärdenfors and Makinson [22] specified postulates for contraction and revision operators, which they claim should be satisfied by all rational belief change operators.

The problems described hold for the change of ontologies as well, as an ontology can be considered a belief base in the sense of the belief change theory. In this context, the problem is known as *ontology change*. Belief change theory can not be directly applied to description logics because it is based on assumptions that generally fail for description logics [23]. However, the authors in [24] show that all logics admit a contraction operator that satisfies the postulates except the recovery postulate. In [25], the authors show that the theory can be applied if the recovery postulate is slightly generalized.

Another approach to the problem of ontology update is taken in [26], which proposes an ontology update framework where ontology update specifications describe certain change patterns that can be performed. Change requests (adding or removing pieces of information to the ontology) are only accepted if the corresponding update specification accounts for it. The update specification is implemented similar to a database trigger and possibly carries out more ontology changes than explicitly requested to ensure ontology consistency.

Updating the ontology can be avoided, when changes over time are modeled in the ontology. For this approach, so-called fluents are used, where facts are tagged with the time or range of time at which they are valid. The authors in [27] describe how fluents can be modeled in OWL. However, this creates a large number of additional instances in the ontology, which makes it impractical for the application in the IT management context, where many changes of the ontology take place in short time frames.

III. Concepts

A. Transformation of the Common Information Model to OWL

As pointed out in Section II-B, a translation of CIM to OWL must be performed. The translation approach described in this section has first been published in [11] and is described in full detail in [28]. This section gives an overview of the translation approach and describes (previously unpublished) technical details necessary for the implementation of the translation using exemplary values. Note that the resulting OWL ontology is available for download at [29].

The translation creates an ontology that consists of two parts. The first part is a manually modeled meta-ontology that describes super classes, properties and annotations that meta-model CIM constructs, which can not be directly translated to OWL. The meta ontology has the namespace `cim-meta`. The second part is the CIM schema ontology, which is modeled using OWL-, RDFS- and CIM meta constructs, and which represents the actual CIM model. This part is generated programmatically by parsing the original MOF files and applying the following translation rules. The implementation uses pattern matching techniques on the abstract syntax tree of the CIM model to apply the translation rules.

Structural translation is mostly straightforward. CIM classes can be mapped to OWL classes, although a class in object-oriented modeling is not identical to the concept of a class in an ontology. Likewise, generalisation (inheritance) can be expressed using the OWL subclass concept `rdfs:subClassOf`. CIM has another basic construct for the expression of relationships, a so-called Association, which is a special kind of class with two typed reference properties (antecedent and dependent). Associations are mapped to OWL classes that inherit from the special meta class `cim-meta:CIM_Association`. CIM aggregations are handled accordingly.

Each CIM property is translated into an OWL object property and an OWL class that inherits from `cim-meta:CIM_Value`. The domain of the object property is the class that

originally contained the property, while the range is the CIM_Value subclass. This class in turn then has a data property, which contains the actual value. The additional indirection is necessary for two reasons: CIM properties can have values of both primitive types such as uint32 and references to classes, also the CIM_Value subclass is necessary to be able to express the CIM qualifiers on the property.

```
1 class CIM_System : CIM_EnabledLogicalElement {
2   string Name;
3 }
4
5 class CIM_ComputerSystem : CIM_System {
6   uint32 SetPowerState(uint32 PowerState, datetime
        Time);
7 }
```

Listing 1. CIM properties

In the following paragraphs, the mapping is illustrated using a concrete example. For the first class in Listing 1 the following OWL elements are created:

- An OWL class CIM_System that is a subclass of the OWL class CIM_EnabledLogicalElement
- An OWL class CIM_System__Name_Value that is a subclass of cim-meta:CIM_Value
- An OWL object property CIM_System__Name with domain CIM_System and range CIM_System__Name_Value

Information about qualifiers on the property can then be added to the CIM_System__Name_Value class as annotations or further object or data properties.

To translate methods, even more structural elements are required. The method itself, its parameters and return type, and the types of each parameter must be modeled. The second class in Listing 1 is translated into the following CIM elements:

- An OWL class CIM_ComputerSystem that is a subclass of the OWL class CIM_System
- An OWL instance CIM_ComputerSystem__SetPowerState_Method that is an instance of cim-meta:CIM_Method (which has data properties cim-meta:methodName and cim-meta:methodType)
- An OWL object property CIM_ComputerSystem__SetPowerState that has the domain CIM_ComputerSystem and the range cim-meta:CIM_Method and an annotation cim-meta:methodInstance that points to the instance
- An OWL object property CIM_ComputerSystem__SetPowerState__Parameters that has the method instance as domain and a range of an owl:oneOf enumeration
- The enumeration contains the instances CIM_ComputerSystem__SetPowerState_Parameters_PowerState and CIM_ComputerSystem__SetPowerState_Parameters_Time, which are both instances of cim-meta:CIM_Method_Parameter, which in turn has the data properties cim-meta:parameterName, cim-meta:parameterType and cim-meta:parameterPosition.

CIM datatypes are translated into the corresponding XSD datatype, as shown in table I.

Table I. Translation of CIM types to XSD types

CIM type	XSD type	CIM type	XSD type
uint8	unsignedByte	string	string
sint8	byte	boolean	boolean
uint16	unsignedShort	real32	float
sint16	short	real64	double
sint32	int	datetime	dateTime
uint32	unsignedInt	char16	string
sint64	long	uint64	unsignedLong

The translation of CIM qualifiers is performed by expressing the semantics of each qualifier using OWL features, as far as possible. In cases where this is not possible, corresponding classes and properties are modeled in the CIM meta ontology that the schema ontology can refer to. Table II gives an overview of the translation of CIM structures and qualifiers to OWL.

Table II. Translation of CIM constructs to OWL

CIM Construct	Translation in OWL
Abstract	cim-meta:isAbstract
Aggregate	Handled together with Aggregation
Aggregation	cim-meta:CIM_Aggregation, cim-meta:CIM_Aggregation_Parent, cim-meta:CIM_Aggregation_Child
Alias	owl:equivalentProperty
Association	cim-meta:CIM_Association, cim-meta:CIM_Association_Role
Class	owl:Class
ClassConstraint	cim-meta:classConstraint
Composition	cim-meta:CIM_Composition, cim-meta:CIM_Composition_Parent, cim-meta:CIM_Composition_Child
Correlatable	No translation
Datatypes	See table I
Default values	Union of original property range and default value singleton
Deprecated	owl:DeprecatedClass, owl:DeprecatedProperty
Description	rdfs:comment
DisplayName	cim-meta:displayName
Exception	cim-meta:exception

Experimental	cim-meta:experimental
In	cim-meta:in
Inheritance	rdfs:subClassOf
Key	owl:inverseFunctionalProperty
MappingStrings	cim-meta:mappingStrings
Max	owl:maxCardinality
MaxLen	owl:Restriction, xsd:maxLength
MaxValue	owl:Restriction, xsd:maxInclusive
Methods	cim-meta:CIM_Method (plus one instance), cim-meta:CIM_Method_Parameter (plus one instance), cim-meta:methodName, cim-meta:methodType, cim-meta:parameterName, cim-meta:parameterType, cim-meta:parameterPosition, object property for method, object property for method parameters
Method-Constraint	cim-meta:methodConstraint
Min	owl:minCardinality
MinLen	owl:Restriction, xsd:minLength
MinValue	owl:Restriction, xsd:minInclusive
ModelCorrespondence	rdfs:seeAlso
Out	cim-meta:out
Override	rdfs:subPropertyOf
Property	cim:CIM_<Class>__<Property>_Value (subclass of cim-meta:CIM_Value), cim:CIM_<Class>__<Property>
Property-Constraint	cim-meta:propertyConstraint
PUnit	cim-meta:punit
UMLPackage-Path	cim-meta:UMLPackagePath
Read	cim-meta:readable
Required	owl:minCardinality of 1
Terminal	cim-meta:isTerminal
Units	cim-meta:units
ValueMap	cim-meta:valueMap
Values	cim-meta:value
Version	owl:versionInfo
Write	cim-meta:writable

B. Combination of Ontologies and Bayesian Networks

The goal of the new architecture is to combine precise and probabilistic knowledge. A central concept is to model an ontology with entities and relationships and to derive the Bayesian Network dynamically from the ontology. Therefore, several specifications need to be defined to put some additional semantics into the ontology. Mainly,

- how random variables are represented in the ontology and
- how relationships are represented in the ontology.

1) Variable Representation: As mentioned before, OWL ontologies are able to represent continuous and discrete variables as data properties. As Bayesian Networks only operate on discrete random variables, a discretization must be applied. To discretize continuous variables, some additional information is needed. OWL does not support the adding of supplemental data to data property assertions. In [5], this problem was solved by adding the data on the RDF layer of OWL. The approach veers away from the concepts of OWL and the support for most editors and tools gets lost. Hence, for this approach another representation is used, which capsulates random variables through instances of variable classes, which has a data property that contains the actual value of the variable. There are three different types of variables:

- Continuous variables
- Discrete variables
- Enumerations

Continuous variables are containers for floating point values, discrete variables are containers for integers. In contrast enumerations do not store primitive data but OWL individuals as a value.

A mechanism is needed to map values of all three types of variables from the ontology to the generated Bayesian Network and back again. Since enumerations generally have just a small state space the values can be mapped one by one. For continuous and discrete variables the mapping is problematic and a discretization must be applied. A discretization for variables that are already discrete is needed, because the state space (in most cases full integer state space) is too large for Bayesian Networks. For a random variable x the size of its conditional probability table $probsize(x)$ grows with

$$probsize(x) = spacesize(x) \cdot \prod_{c \in cause(x)} spacesize(c) \quad (1)$$

where $cause(x)$ is the set of causing variables of variable x and $spacesize(c)$ is the size of the state space of the variable c.

To support the discretization mechanism, we defined special interval properties for both types of variables. These intervals are used to discretize values in the runtime environment. They are defined in a math-based syntax as additional data properties of the variable class. For the mapping from OWL to the network, the matching interval is taken. Every interval is an unique discrete value in the network. The mapping back from the network to OWL is more complicated, because the discrete data has to be enriched. For this, the median of the matching interval is used.

2) Relationship Representation: Another part of the OWL model are relationships. They describe the coherence be-

tween random variables of the system. Three types of relationships are considered in the model:

- Functional relationships
- Causal relationships
- Correlations

Functional relationships are based on known dependencies, which can be expressed by a mathematical formula. SWRL as part of the OWL specification already supports the usage of mathematical expressions and is used for the definition of functional relationships. The rules are evaluated at runtime by an OWL reasoner and new axioms are generated depending on the bound variables. Because correlations can be seen as bidirectional edges and causal relationships as unidirectional edges, the OWL object property concept can be used for the representation of both. In general it is not possible to connect data properties in OWL, but in this case it is feasible because all variables are already encapsulated by instances of the variable class.

3) Ontology Structure: A base ontology has been defined, which defined all OWL concepts needed for the use of the specialized variables and relationships in a client-specific ontology. Figure 2 shows the structure of the ontology, which is described in detail as follows.

- The `Variable` class is the base class for all variable types. It is the domain and the range for the `causation` object property, which defines causation between variables and the `correlation` object property that defines a correlation between variables.
- The `NumericVariable` class extends the variable class and is the base class for all numerical variables. It is the domain of the `interval` data property, which defines intervals for discretization and the the `unit` data property that defines the unit of a numerical variable.
- The `ContinuousVariable` class extends the numerical variable class and is used for continuous variables in the model. It is the domain of a `continuousValue` data property, which stores the discrete value of a continuous variable.
- The `DiscreteVariable` class extends the numerical variable class and is used for discrete variables in the model. It is the domain of a `discreteValue` data property, which stores the discrete value of a discrete variable.
- The `Enumeration` class extends the variable class and is used for enumerations in the model. It is the domain of the `enumerationClass` data property, which defines an OWL class as state space and the `enumerationValue` object property that stores the value of an enumeration variable.
- The `System` class, which is the base class for all root nodes of defined systems.

The ontology can be imported into any other ontology and instances of the classes and assertions of the properties can be created. To create a system model independent of these special properties and classes but capable of using the features it is possible to import both the system ontology and the variable and relationships defining ontology to a new ontology and define same individual, same object property and same data property axioms.

4) Joint Model Evaluation: For the evaluation of the relationships between variables different techniques are used. Since functional relationships are already defined as SWRL rules, the evaluation is simple. An OWL reasoner binds the variables in the body of each rule and creates the axioms in the head of the rule.

Causations are mapped to a Bayesian Network where each instance of the variable class becomes a node. For numerical variables each variable is checked for intervals. A discrete state is created for each interval in the state space of the node in the network. Enumerations are checked for their defined enumeration class and for each individual of this class a state is created with the unique name of the individual. Causal relationships between variables become arcs in the Bayesian Network.

After the structure of the network has been created the runtime values are mapped from the individuals in the ontology to the nodes in the network in every reasoning cycle. For a numerical variable the value is read, the fitting interval is found and the state of the node is set to the unique interval. For enumerations the individual is extracted and the state of the node in the network is set to the unique identifier of the individual.

In the next step an inference algorithm is applied to calculate the belief for the states of each unobserved variable (variables which have no value set in the ontology) which part a causal relationship. If the calculated belief is above a defined threshold the deduced value is fed back into the ontology as an property assertion for the variable. For enumerations that step is quite simple because the state is exactly the identifier of the individual. In case of numerical variables the matching interval is found and the median of the interval is set as value for the variable in the ontology. Values derived from the Bayesian Network are marked as being fuzzy by a property in the ontology so that other inference algorithms are aware of that fact.

Correlations are more complex to handle because they are based on precise knowledge (that the correlation exists) but the coherence between the variables is just a statistical measurement. An evaluation compared to that of causations is not possible, because correlations are bidirectional and thus cannot be presented in an acyclic graph. Therefore, correlations are analyzed offline for their functional relationship and are replaced by SWRL rules. These rules can be evaluated like the rules for functional relationships, the result is marked as being fuzzy as well.

Given that system facts can be revoked a mechanism is needed to revoke facts in the ontology as well as facts, which

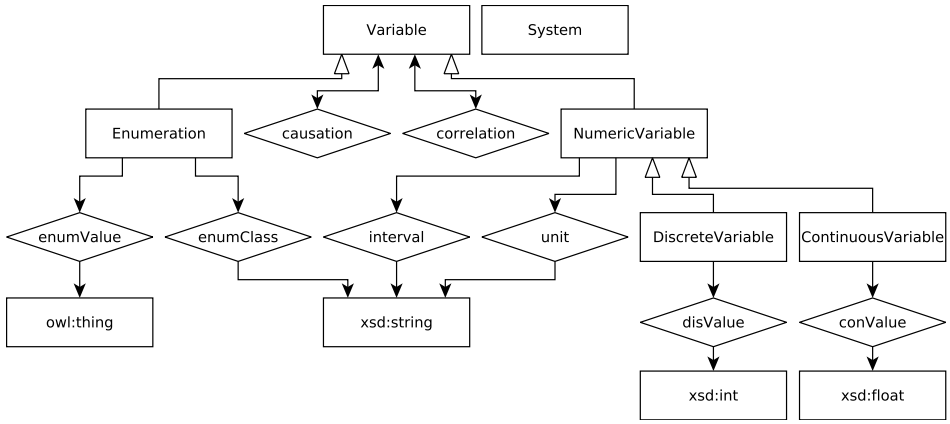


Figure 2. Entity structure of the created ontology.

have been derived based on those facts. OWL does not have a concept to automatically remove depending facts. Hence a directed acyclic graph is used to store the dependencies between the axioms of the ontology. If an axiom is removed, all child axioms are removed as well. Thus, not the whole derived knowledge must be revoked but only a subset of axioms.

IV. Architecture

A new architecture for ontology-based automated IT management is currently under development by the authors and the main ideas are sketched in this section. The architecture consists of a set of components (shown in Figure 3), which can be grouped into

- Importers that add new data to the ontology
- Reasoning components, which use the existing data to derive new knowledge
- Management components, which interact with the system under management.

The central element of the system is an ontology that is used as a shared knowledge base (blackboard) for all components. Each component can read data from the knowledge base and add or remove facts from it. Service invocations are used for the inter-component communication. The architecture is designed to be used in a distributed fashion.

A. Importers

The combination of different domain models raises the requirement for corresponding importers. These specific components know how to map the domain specific model to an ontology model. Hence, an interface is defined, which allows the use of new domain specific model importers. Implemented model importers are an ontology importer and a CIM importer. The ontology importer simply reads the data from an OWL ontology and adds the facts to the shared knowledge base. The CIM importer uses the mapping rules described in III-A to map the CIM schema to OWL facts.

As well as models, rules can be specified in a domain specific manner. Hence, an interface is provided for the implementation of domain specific rule importers. Internally, SWRL is used as rule format for the shared ontology and an according importer was implemented.

In general, the domain model contains just the taxonomy of the monitored system but not the instance data. Therefore, a component is needed that monitors the system under management and imports runtime data into the ontology by creating according instances. Such components are called instance importers. An interface is provided for the integration of domain specific instance importers. Already implemented instance importers are the log record importer, which maps log records to instances and relations, and the CIM instance importer, which uses the OpenPegasus CIMOM to get information from a CIM-based management system. Other application-specific instance importers can be added as needed.

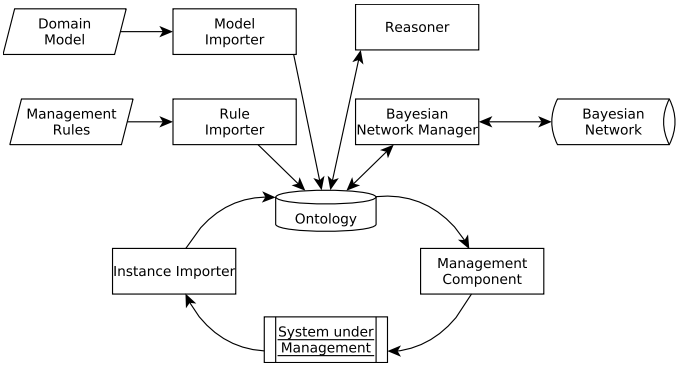


Figure 3. Components of the developed architecture

B. Reasoning

The strength of OWL and its formal grounding is the ability to reason new knowledge from an existing knowledge base. In our architecture this feature is used to derive new

facts from the domain specific models, the imported rules and the monitored instance data.

In many cases it is insufficient to just consider exact knowledge in IT management, because side effects and complex relationships are either not known or can not be modeled in an adequate level of abstraction. But especially for state prediction and root cause analysis probabilistic knowledge and the statistical consideration of historical data are needed. Because of that, a concept is used to make probabilistic modeling and reasoning possible, which is described in detail in III-B. The structure of the Bayesian network is derived from the OWL model. The conditional probabilities are not modeled in the ontology directly, but trained using a maximum likelihood algorithm during a precedent training phase, which uses real data from the system under management.

In the next step the OWL model is analyzed for variable states, which will be set as evidences in the Bayesian network. Subsequently, an inference algorithm is applied to calculate the belief for the states of unobserved variables (variables which have no value set in the ontology). If the calculated belief is above a defined threshold, the deduced value is set for the variable in the ontology and can thereby be used by the exact reasoners for further reasoning. To ensure the knowledge exchange between the reasoning components a component can be called multiple times in a reasoning cycle.

C. Management components

Management components are used to reconfigure the system under management. They contain the knowledge that is needed to interact with a specific component of the system. Depending on the evaluation results of the rules, according actions are triggered. When CIM is used as a domain model, the management components can call methods on the CIMOM, which in turn controls the particular component, or it can execute external commands directly.

D. Runtime

The first step on application startup is the import of required domain models and rules using the according model and rule importers. After that, the management cycle is started (also known as MAPE-K loop [30], which stands for monitor, analyze, plan, execute and knowledge). The loop begins with the monitoring phase, where information from the system under management is read and imported into the ontology as instances.

In the analysis phase, the domain models, the rules and the monitored data are used for the reasoning of new knowledge. The reasoning process is shown in Figure 4.

The base ontology contains all the imported and monitored data. When the reasoning process starts, all data of the base ontology is copied into the working ontology. All reasoners are applied to this ontology sequentially and add their reasoned knowledge to it. When all reasoners have finished, the data of the working ontology is copied to the reasoned ontology, which is used for queries into the knowledge base and stays untouched until the next reasoning phase has finished.

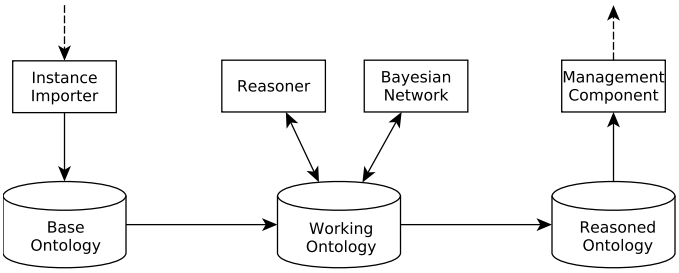


Figure 4. Multi step ontology reasoning process

The reasoning takes place in this multi-step process for two reasons: The first reason is handling ontology change, as new information can be added easily to an ontology, but not retracted easily. By keeping the base model and inferred knowledge from different reasoners in separate sub-ontologies, inferred knowledge from a single reasoner can be retracted without effort. The second reason is that the last version of the *reasoned ontology* can still be queried, while the new version is being created. As reasoning can be slow on large ontologies, this makes sure that clients do not block on queries but can always receive an instant reply. The query result therefore may be as old as one reasoning cycle.

The last steps in the cycle are the plan and execute phases. The management components use the data of the reasoned ontology to make management decisions and execute them on the system under management. The presented architecture was prototypically implemented in Java using the OSGi Framework as service middleware. The implementation comprises the core components and the specific features described above (i.e., model importers, rule importers and instance importers). More domain-specific components will be added in the course further application of the approach (see Section V).

For the service abstraction the interfaces of the OWL API are used.

V. Conclusion and Future Work

In this paper we presented an approach for ontology-based IT management. The approach comprises an architecture that uses an ontology which integrates the domain model, rules and dynamically updated instance data. Two main problems were solved: The first problem is the creation of a suitable domain model, which was covered by the translation of CIM to OWL and the expression of probabilistic knowledge using Bayesian networks. The integration of other domain models has yet to be examined. The second problem is the

continuous update of the ontology with new facts. This is a topic of current research, and our solution is a multi-step reasoning process. Performance comparisons to other approaches and with different ontologies must be conducted.

Future work includes the development of importers for other domain models. The application of the presented approach is currently underway in two different concrete domains: One is the domain of storage management, the other is the domain of ambient assisted living (AAL). The application in these domains includes the development of domain-specific importers and the overall optimization of performance of the runtime system.

In the context of storage management the Storage Management Initiative Specification (SMI-S), which is a specialization of the CIM Model, can be used to manage storage systems. Rules, which are verbally defined in the specification, are formalized and integrated into the OWL model. Besides, the probabilistic part is used to make assertions about future states (e.g., how high is the probability of a full file system tomorrow if there is a peak) and to analyze previous scenarios (e.g., what was the most likely reason for a file server crash). In combination a pro-active management can be achieved and systems can be reconfigured before a failure occurs.

In the context of ambient assisted living the domain is a living environment, equipped with a set of sensors and effectors. That environment is modeled in a hierarchy of ontologies and monitored during runtime. The observed data is used to derive higher level knowledge, e.g., that lights should automatically be switched on or off when a person enters a room.

The proof-of-concept implementation of the ontology-based management system and the integration of probabilistic knowledge with the OWL ontology enables rule-based automatic management of domains for which a domain ontology was created.

References

[1] A. Textor, F. Meyer, and R. Kroeger, "Semantic Processing in IT Management," in *Proceedings of the Fifth International Conference on Advances in Semantic Processing (SEMAPRO)*, Lisbon, Portugal, November 2011.

[2] Distributed Management Task Force, "Common Information Model (CIM)," http://www.dmtf.org/standards/cim/, 2012-12-18.

[3] World Wide Web Consortium, "OWL Web Ontology Language," http://www.w3.org/TR/owl2-overview/, 2012-12-18.

[4] Object Management Group, "Unified Modeling Language (UML)," http://uml.org/, 2012-12-18.

[5] J. E. L. De Vergara, V. A. Villagrá, and J. Berrocal, "Applying the Web ontology language to management information definitions," *IEEE Communications Magazine*, vol. 42, no. 7, pp. 68–74, July 2004.

[6] J. E. L. De Vergara, A. Guerrero, V. A. Villagrá, and J. Berrocal, "Ontology-Based Network Management: Study Cases and Lessons Learned," *Journal of Network and Systems Management*, vol. 17, no. 3, pp. 234–254, September 2009.

[7] A. Guerrero, V. A. Villagrá, J. E. L. de Vergara, A. Sánchez-Macián, and J. Berrocal, "Ontology-Based Policy Refinement Using SWRL Rules for Management Information Definitions in OWL," *Large Scale Management of Distributed Systems*, vol. 4269, pp. 227–232, October 2006.

[8] A. Westerinen, J. Schnizlein, J. Strassner, M. Scherling, B. Quinn, S. Herzog, A. Huynh, M. Carlson, J. Perry, and S. Waldbusser, "Terminology for policy-based management." The Internet Society, November 2001.

[9] S. Quirolgico, P. Assis, A. Westerinen, M. Baskey, and E. Stokes, "Toward a Formal Common Information Model Ontology," pp. 11–21, November 2004.

[10] M. Majewska, B. Kryza, and J. Kitowski, *Translation of Common Information Model to Web Ontology Language*, ser. Lecture Notes in Computer Science. Berlin: Springer Berlin Heidelberg, May 2007, vol. 4487, pp. 414–417.

[11] A. Textor, J. Stynes, and R. Kroeger, "Transformation of the Common Information Model to OWL," in *10th International Conference on Web Engineering - ICWE 2010 Workshops*, ser. LNCS, vol. 6385. Springer Verlag, July 2010, pp. 163–174.

[12] G. Hamerly and C. Elkan, "Bayesian approaches to failure prediction for disk drives," in *In Proceedings of the eighteenth international conference on machine learning*. Morgan Kaufmann, July 2001, pp. 202–209.

[13] T. M. Mitchell, *Machine Learning*, 1st ed. McGraw-Hill Science/Engineering/Math, March 1997.

[14] O. Chapelle, B. Schölkopf, and A. Zien, Eds., *Semi-Supervised Learning (Adaptive Computation and Machine Learning)*. The MIT Press, September 2006.

[15] C.-G. Bai, "Bayesian network based software reliability prediction with an operational profile," *J. Syst. Softw.*, vol. 77, no. 2, pp. 103–112, August 2005.

[16] J. D. Musa, "Operational profiles in software-reliability engineering," *IEEE Software*, vol. 10, pp. 14–32, March 1993.

[17] M. Kaaniche and K. Kanoun, "The discrete time hyperexponential model for software reliability growth evaluation," in *Software Reliability Engineering, 1992. Proceedings., Third International Symposium on*, October 1992, pp. 64 –75.

[18] Y. Liu, W. P. Cheah, B.-K. Kim, and H. Park, "Predict software failure-prone by learning bayesian network," *International Journal of Advanced Science and Technology*, vol. 35, 2006.

[19] Eclipse Foundation, "The Eclipse Project," http://www.eclipse.org/, 2012-12-18.

[20] Z. Ding and Y. Peng, "A probabilistic extension to ontology language owl," in *In Proceedings of the 37th Hawaii International Conference On System Sciences (HICSS-37), Big Island*, January 2004.

[21] P. Cesar, G. Costa, K. B. Laskey, and K. J. Laskey, "Pr-owl: A bayesian ontology language for the semantic web," in *Center for Technology-Enhanced Learning, University of Missouri-Rolla*, 2003.

[22] C. E. Alchourrón, P. Gärdenfors, and D. Makinson, "On the logic of theory change: Partial meet contraction and revision functions," *The Journal of Symbolic Logic*, vol. 50, pp. 510 – 530, June 1985.

[23] G. Qi and F. Yang, "A survey of revision approaches in description logics," in *Web Reasoning and Rule Systems*, ser. LNCS, vol. 5341. Springer, November 2008, pp. 74–88.

[24] G. Flouris, D. Plexousakis, and G. Antoniou, "AGM Postulates in Arbitrary Logics: Initial Results and Applications," 2004.

[25] M. M. Ribeiro and R. Wassermann, "First steps towards revising ontologies," in *Proc. of WONRO'2006*, 2006.

[26] U. Lösch, S. Rudolph, D. Vrandečić, and R. Studer, "Tempus fugit - towards an ontology update language," in *6th European Semantic Web Conference (ESWC 09)*, vol. 1. Springer, January 2009, pp. 278–292.

[27] C. Welty, R. Fikes, S. Makarios, C. Welty, R. Fikes, and S. Makarios, "A reusable ontology for fluents in owl," in *In Proceedings of Formal Ontology in Information Systems (FOIS)*, November 2006, pp. 226–236.

[28] A. Textor, "Semi-Automatic Management of Knowledge Bases Using Formal ontologies," Master's thesis, Cork Institute of Technology, Ireland, March 2011.

[29] RheinMain University of Applied Sciences, Distributed Systems Lab, "CIM OWL Ontology," http://wwwvs.cs.hs-rm.de/oss/cimowl/index.html, 2012-12-18.

[30] IBM Corporation, "An Architectural Blueprint for Autonomic Computing, Technical Whitepaper (Fourth Edition)," June 2006.

Parallel SPARQL Query Processing Using Bobox

Zbyněk Falt, Miroslav Čermák, Jiří Dokulil, and Filip Zavoral
Charles University in Prague, Czech Republic
{*falt,cermak,dokulil,zavoral*}*@ksi.mff.cuni.cz*

***Abstract*—Proliferation of RDF data on the Web creates a need for systems that are not only capable of querying them, but also capable of scaling efficiently with the growing size of the data. Parallelization is one of the ways of achieving this goal. There is also room for optimization in RDF processing to reduce the gap between RDF and relational data processing. SPARQL is a popular RDF query language; however current engines do not fully benefit from parallelization potential. We present a solution that makes use of the Bobox platform, which was designed to support development of data-intensive parallel computations as a powerful tool for querying RDF data stores. A key part of the solution is a SPARQL compiler and execution plan optimizer, which were tailored specifically to work with the Bobox parallel framework. The experiments described in this paper show that such a parallel approach to RDF data processing has a potential to provide better performance than current serial engines.**

Keywords-SPARQL; Bobox; query optimization; parallel.

I. Introduction

SPARQL [2] is a query language for RDF [3] (Resource Definition Framework) widely used in semantic web databases. It contains capabilities for querying graph patterns along with their conjunctions and disjunctions. SPARQL algebra is similar to relational algebra; however, there are several important differences, such as the absence of NULL values. As a result of these differences, the application of relational algebra into semantic processing is not straightforward and the algorithms have to be adapted so it is possible to use them.

As the prevalence of semantic data on the web is getting bigger, the Semantic Web databases are growing in size. There are two main approaches to storing and accessing these data efficiently: using traditional relational means or using semantic tools, such as different RDF triplestores [3] accessed using SPARQL. Semantic tools are still in development and a lot of effort is given to the research of effective storing of RDF data and their querying [4]. One way of improving performance is the use of modern, multicore CPUs in parallel processing.

Nowadays, there are several database engines which are capable of evaluating SPARQL queries, such as SESAME [5], JENA [6], Virtuoso [7], OWLIM [8] or RDF-3X [9], that is currently considered to be one of the fastest single node RDF-store [10]. These stores support parallel computation of multiple queries; however, they mostly do not use the potential of parallel computation of particular queries.

The Bobox framework [11], [12], [13] was designed to support the development of data-intensive parallel computations. The main idea behind Bobox is to divide a large task into many simple tasks that can be arranged into a non-linear pipeline. The tasks are executed in parallel and the execution is driven by the availability of data on their inputs. The developer does not have to be concerned about problems such as synchronization, scheduling and race conditions. All this is done by the framework. The system can be easily used as a database execution engine; however, each query language requires its own front-end that translates a request (query) into a definition of the structure of the pipeline that corresponds to the query.

In the paper, we present a tool for efficient parallel querying of RDF data [14] using SPARQL build on top of the Bobox framework [1], [15]. The data are stored using an in-memory triple store. We provide a description of query processing using SPARQL-specific parts of the Bobox and provide results of benchmarks. Benchmarks were performed using the SP^{2}Bench [16] query set and data generator.

The rest of the paper is structured as follows: Section II describes the Bobox framework. Models used to represent queries and a description of query processing is contained in Section III. Data representation and the implementation of operators using Bobox framework is described in Section IV. Section V presents our experiments and a discussion of their results. Section VI compares our solution to other contemporary parallelization frameworks. Section VII describes future research directions and concludes the paper.

II. Bobox Framework

A. Bobox Architecture

Bobox is a parallelization framework which simplifies writing parallel, data intensive programs and serves as a testbed for the development of generic and especially data-oriented parallel algorithms.

Bobox provides a run-time environment which is used to execute a non-linear pipeline (we denote it as the *execution plan*) in parallel. The execution plan consists of computational units (we denote them as the *boxes*) which are connected together by directed edges. The task of each box is to receive data from its incoming edges (i.e. from its *inputs*) and to send the resulting data to its outgoing edges

(i.e. to its *outputs*). The user provides the execution plan (i.e. the implementation of boxes and their mutual connections) and passes it to the framework which is responsible for the evaluation of the plan.

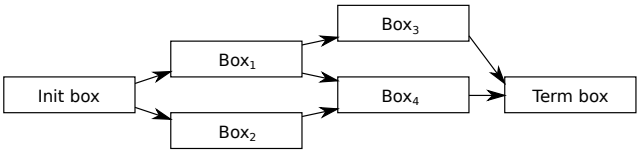


Figure 1. Example of an execution plan

Figure 1 shows an example of an execution plan. Each plan must contain two special boxes:

- *init box* – this is the first box (in a topological order) of the plan which is executed.
- *term box* – this is the last box and denotes that the execution plan was completely evaluated.

The implementation of boxes is quite straightforward and simple, since Bobox provides a very powerful and easy to use interface for their development. Additionally, the source code is expected to be strictly single-threaded. Therefore, the developer does not have to be familiar with parallel programming. Although this requirement on the source code may seam limiting, the framework is especially targeted to a development of highly scalable applications [17].

The only communication between boxes is done by sending *envelopes* (communication units containing data) along their outgoing edges. Each envelope consists of several columns and each column contains a certain number of data items. The data type of items in one column must be the same in all envelopes transferred along one particular edge; however, different columns in one envelope may have different data types. The data types of these columns are defined by the execution plan.

The number of data items in all columns in one envelope must be always the same. Therefore, we may define the list of i-th items of all columns in one envelope as its i-th *data line*. The Figure 2 shows an example of an envelope.

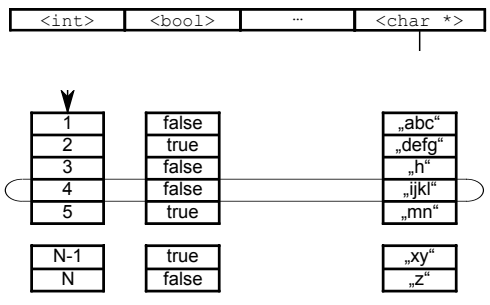


Figure 2. The structure of an envelope

The total number of data lines in an envelope is chosen according to the size of cache memories in the system. Therefore, the communication may take place completely in cache memory. This increases the efficiency of processing of incoming envelopes by a box.

Currently, only shared-memory architectures are supported; therefore, the only shared pointers to the envelopes are transferred. This speeds up operations such as broadcast box (i.e., the box which resends its input to its outputs) significantly since they do not have to access data stored in envelopes.

There is one special envelope (so called *poisoned pill*) which is sent after the last regular envelope to close the output of a source box. For the receiver of the poisoned pill it is a signal that all data were already received on that input.

In fact, the only work which is done by the init box is sending the poisoned pill to its output and the only responsibility of the term box is to terminate the evaluation of the execution plan when it receives the poisoned pill on its input.

The interface of Bobox for box development is very flexible; therefore, the developer of a box may choose between multiple views on the data communication:

- The communication is a stream of envelopes. This is useful for efficient implementation of boxes which do not have to access data in envelopes such as broadcast box or stream splitter (see Section IV-D) or implementation of boxes which process their inputs by envelopes.
- The communication is a stream of data lines. This is useful for easier implementation of boxes which manipulate with data lines one by one such as filter box (see Section IV-C2).
- The combination of both views. For example, the sort box (see IV-C3) processes input by envelopes, but produces the output as a stream of data lines.

Although the body of boxes must be strictly single-threaded, Bobox may introduce three types of parallelism:

1) Task parallelism, when independent streams are processed in parallel.
2) Pipeline parallelism, when the producer of a stream runs in parallel with its consumer.
3) Data parallelism, when independent parts of one streams are processed in parallel.

The first two types of parallelism are exploited implicitly during the evaluation of a plan. Therefore, even an application which does not contain any explicit parallelism may benefit from multiple processors in the system (see Section V-A). Data parallelism must be explicitly stated in the execution plan by the user (see IV-D); however, it is still much easier to modify the execution plan than writing parallel code by hand.

B. Flow control

Each box has only limited buffer for incoming envelopes. When this buffer becomes full, the producer of the envelopes

is suspended until at least one envelope from the buffer is processed. This strategy increases the performance of the system since the operators which produce data faster than their consumers are able to process are suspended to not to consume the CPU time uselessly. This time may be used to execute other boxes. Additionally, this method yields to lower memory consumption, since there is only a limited number of unprocessed slots which occupy the memory at a time.

On the other hand, this flow control may sometimes yield to a deadlock (see Section V-C) or may limit the level of parallelism (see Section IV-C6 for an example).

C. Box scheduling

Scheduling of boxes is a very important factor which significantly influences the performance of Bobox. The scheduling strategies are described in a more detail in [12].

During the initialization of Bobox, a same number of worker threads as the number of physical processors is created. Only these worker threads may execute the code of boxes. The scheduler has two main data structures:

- Each worker thread has its own double ended queue of *immediate tasks*.
- Each execution plan which is being evaluated has its own queue of *deferred tasks*.

There are three cases when a box is scheduled:

- When a new execution plan is about to evaluate, a new queue of deferred tasks for that plan is created and its init box is put to the front of that queue.
- When a box sends an envelope to another box, the destination box is put to the front of the queue of immediate tasks of the thread which is executing the source box.
- When a box stops to be suspended because of flow control, it is put to the queue of deferred tasks of the corresponding plan.

When the working thread is ready to execute a box, it choose the first existing box in this order:

1) The newest box in its queue of immediate tasks. This box receives an envelope created by this thread recently. Therefore, it is probable that this envelope is completely hot in a cache so accessing its data is probably much faster than accessing other envelopes.
2) The oldest box in the queue of deferred tasks of the oldest execution plan. This ensures that scheduling of deferred tasks of one execution plan are scheduled fairly. However, the execution plans are prioritized according to their age – the older the execution plan is, the higher priority it has. Each evaluation of an execution plan needs some resources (such as memory for envelopes); therefore, the more plans are being evaluated at a time, the more resources are needed for them. This strategy ensures that if there is a box to execute from plans which are currently being executed, no new evaluation is started.
3) The oldest box in the queue of immediate tasks of another worker thread. Worker threads with shared cache memory are prioritized. This avoids suspending of a worker thread despite the fact that there are boxes to execute. Moreover, the oldest box has the lowest probability to have its input hot in a cache memory of the thread from which the box was stolen. Therefore, *stealing* this box should introduce less performance penalty than stealing the newest box in the same queue.

If there is no box to execute, the worker thread is suspended until some other box is scheduled.

Besides the SPARQL compiler described in this paper, the Bobox framework is used in several related projects - model visualization [18], semantic processing [19], [20], query optimization [21], and scheduling in data stream processing [12], [22].

III. Query Representation and Processing

One of the first Bobox applications was SPARQL query evaluator [19]. Since running queries in Bobox needs an appropriate execution plans, SPARQL compiler for Bobox was implemented to generate them from the SPARQL code.

During query processing, the SPARQL compiler uses specialized representation of the query. In the following sections, we mention models used during query rewriting and generation of execution plan.

A. Query Models

Pirahesh et al. [23] proposed the Query Graph Model (QGM) to represent SQL queries. Hartig and Reese [24] modified this model to represent SPARQL queries (SQGM). With appropriate definition of the operations, this model can be easily transformed into a Bobox pipeline definition, so it was an ideal candidate to use.

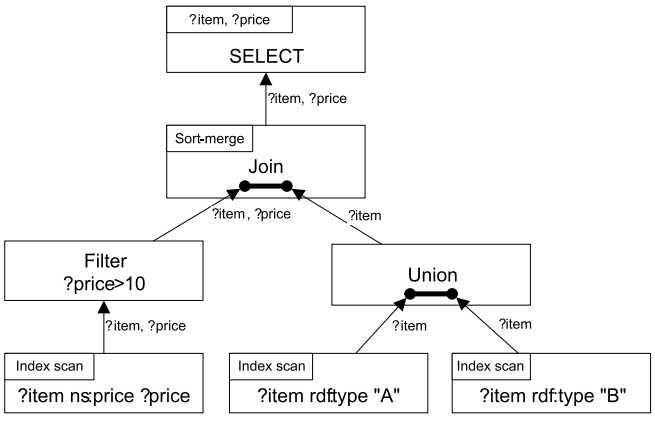


Figure 3. Example of SQGM model.

SQGM model can be interpreted as a directed graph (a directed tree in our case). Nodes represent operators and are depicted as boxes containing headers, body and annotations. Edges represent data flow and are depicted as arrows that follow the direction of the data. Figure 3 shows an example of a simple query represented in the SQGM model. This model is created during an execution plan generation and is used as a definition for the Bobox pipeline.

In [25], we proposed the SPARQL Query Graph Pattern Model (SQGPM) as the model that represents query during optimization steps. This model is focused on representation of the SPARQL query graph patterns [2] rather than on the operations themselves as in the SQGM. It is used to describe relations between group graph patterns (graph patterns consisting of other simple or group graph patterns). The ordering among the graph patterns inside a group graph pattern (or where it is not necessary in order to preserve query equivalency) is undefined. An example of the SQGPM model graphical representation is shown in Figure 4.

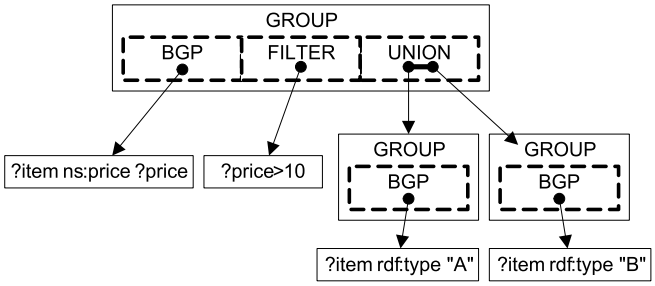


Figure 4. Example of SQGPM model.

Each node in the model represents one group graph pattern that contains an unordered list of references to graph patterns. If the referenced graph pattern is a group graph pattern then it is represented as another SQGPM node. Otherwise the graph pattern is represented by a leaf.

The SQGPM model is built during the syntactical analysis and is modified during the query rewriting step. It is also used as a source model during building the SQGM model.

B. Query Processing

Query processing is performed in a few steps by separate modules of the application as shown in Figure 5. The first steps are performed by the SPARQL front-end represented by compiler. The main goal of these steps is to validate the compiled query, pre-process it and prepare the optimal execution plan according to several heuristics. Execution itself is generated by the Bobox back-end where execution pipeline is initialized according to the plan from the front-end. Following sections describe steps done by the compiler in a more detail way.

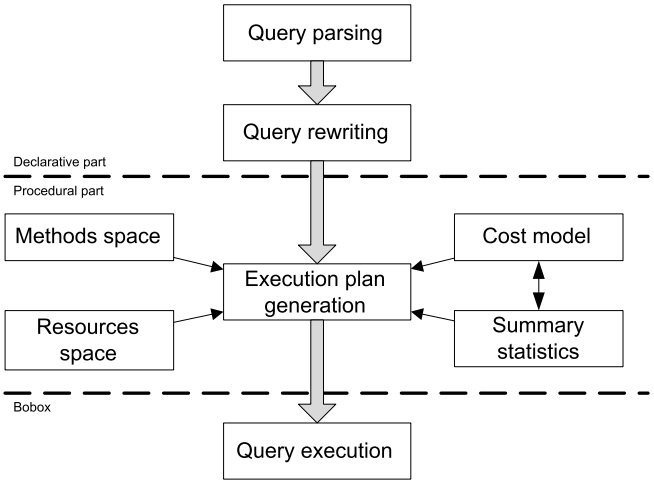


Figure 5. Query processing scheme.

C. Query Parsing and Rewriting

The query parsing step uses standard methods to perform syntactic and lexical analysis according to the W3C recommendation. The input stream is transformed into a SQGPM model. The transformation also includes expanding short forms in queries, replacing aliases and a transformation of blank nodes into variables.

The second step is query rewriting. We cannot expect that all queries are written optimally; they may contain duplicities, constant expressions, inefficient conditions, redundancies, etc. Therefore, the goal of this phase is to normalize queries to achieve a better final performance. We use the following operations:

- Merging of nested *Group graph patterns*
- Duplicities removal
- *Filter*, *Distinct* and *Reduced* propagation
- Projection of variables

During this step, it is necessary to check applicability of each operation with regards to the SPARQL semantics before it is used to preserve query equivalency [25].

D. Execution Plan Generation

In the previous steps, we described some query transformations that resulted in a SQGPM model. However, this model does not specify a complete order of all operations. The main goal of the execution plan generation step is to transform the SQGPM model into an execution plan. This includes selecting orderings of join operations, join types and the best strategy to access the data stored in the physical store.

The query execution plan (e.g., the execution plan of query q5a is depicted in Figure 6) is built from the bottom to the top using dynamic programming to search part of the search space of all possible joins. This strategy is applied to each group graph pattern separately because the order of

the patterns is fixed in the SQGPM model. Also, the result ordering is considered, because a partial plan that seems to be worse locally, but produces a useful ordering of the result, may provide a better overall plan. The list of available atomic operations (e.g., the different types of joins) and their properties are provided by the *Methods Space* module.

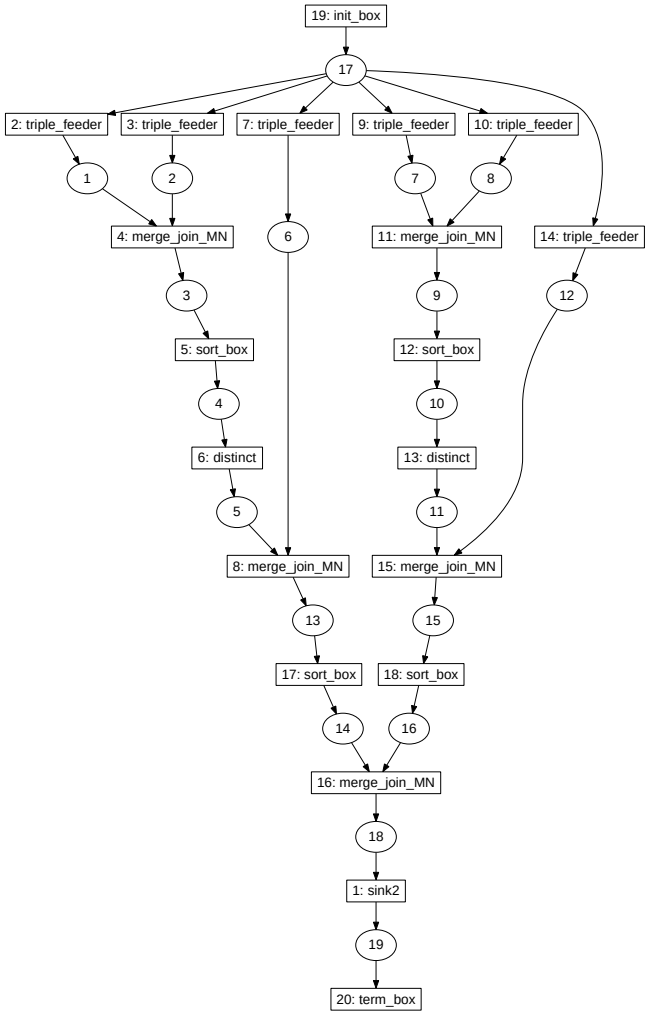


Figure 6. Query execution plan q5a.

In order to compare two execution plans, it is necessary to estimate the *cost* of both plans – an abstract value that represents the projected cost of execution of a plan using the actual data. This is done with the help of the *cost model* that holds information about atomic operation efficiency and *summary statistics* gathered about the stored RDF data.

The search space of all execution plans could be extremely large; we used heuristics to reduce the complexity of the search. Only left-deep trees of join operations are considered. This means that right operand of a join operation may not be another join operation. There is one exception to this rule – avoiding cartesian products. If there is no other way to add another join operation without creating cartesian product, the rest of unused operations is used to build separate tree recursively (using the same algorithm) and the result is joined with the already built tree. This modification greatly improves plans for some of the queries we have tested and often significantly reduces the depth of the tree.

The final execution plan is represented using SQGM model which is serialized into a textual form and passed to the Bobox framework for evaluation.

IV. Evaluation of SPAQRL queries using Bobox

When the compiler finishes the compilation, a query execution plan is generated. This plan must be transformed into a Bobox execution plan and then passed to Bobox for its evaluation. This basically means that operators must be replaced by boxes and they should be connected to form a pipe. Additionally, an efficient representation of data exchanged by boxes must be chosen to process the query efficiently.

A. Data representation

1) Representation of RDF terms: RDF data are typically very redundant, since they contain many duplicities. Many triples typically share the same subjects or predicates. To reduce the number of memory needed for storing the RDF data, we keep only one instance of every unique string and only one instance of every unique term in a memory. Besides the fact that this representation saves the memory, we may represent each term unambiguously by its address. Therefore, for example in case of a join operation, we can test equality of two terms just by a comparison of their addresses.

Additionally, if we need to access the content of a term (e.g. for evaluation of a filter condition) the address can be easily dereferenced. This is faster than the representation of terms by other unique identifiers which would have to be translated to the term in a more complicated way.

2) Representation of RDF database: The database consists of a set of triples. We represent this set as three parallel arrays with the same size which contain addresses of terms in the database. In fact, we keep six copies of these arrays sorted in all possible orders – SPO, SOP, OPS, OSP, PSO and POS. This representation makes implementation of index scans extremely efficient (see IV-C1).

3) Format of envelopes: The format of envelopes is now obvious. It contains columns which correspond to a subset of variables in the query in a form of an address of a particular RDF term. One data line of an envelope corresponds to one possible mapping of variables to their values.

B. Transformation of query execution plan

The output of the compiler is produced completely in a textual form. Therefore, the Bobox must deserialize the query plan first. Despite the fact that this serialization and

deserialization have some overhead, we chose it because of these benefits:

- When distributed computation support is added, the text representation is safer than a binary representation where problems with different formats, encodings or reference types may appear.
- The serialization language has a very simple and effective syntax; serialization and deserialization are much faster than (e.g.) the use of XML. Therefore, the overhead is not so significant.
- The text representation is independent on the programming language; new compilers can be implemented in a different language.
- Compilers can generate plans that contain boxes that have not yet been implemented, which allows earlier testing of the compiler during the development process.
- The query plan may be easily visualized to check the correctness of the compiler. Moreover, the plan might be written by hand which makes the testing of boxes easier. Altogether, this enables debugging of a compiler and Bobox independently on each other.

When the plan is deserialized, the operators in the query execution plan must be replaced by boxes and connected together. The straightforward approach is that each operator in the query execution plan is implemented by exactly one box. Even this approach yields to a parallel evaluation of the plan since pipeline parallelism and task parallelism might be exploited (the query plan has typically a form of a rooted tree with several independent branches). However, it is still usually insufficient to utilize all physical threads available and the most time consuming operations such as nested loops join becomes a bottleneck of the plan. Therefore, they have to be parallelized explicitly. We describe this modification in IV-D.

C. Implementation of query plan operators

1) Index scan: The main objective of a scan operation is to fetch all triples from the database that match the input pattern. Since we keep all triples in all possible orders, it is easy for any input pattern to find the range where all triples which match the pattern are. To find this range, we use binary search. To avoid copying triples from the database to the envelopes, we use the fact that they are stored in parallel arrays. Therefore, we may use the appropriate subarrays directly as columns of output envelopes without data copying.

2) Filter: A filter operation can be implemented in Bobox very easily. The box reads the input as a stream of data lines, evaluates the filter condition on each line and sends out the stream of that data lines which meet the condition.

The evaluation of the filter condition is straightforward since each data line contains addresses of respective RDF terms and by dereferencing them it gets full info about the term such as its type, string/numeric value etc.

3) Sort: Sort is a blocking operation, i.e. it must wait until all input data are received before it starts to produce output data. To increase the pipeline parallelism, we implemented two phase sorting algorithm [17] inspired by external merge sort.

In the first phase, every incoming envelope is sorted independently on other envelopes. This phase is able to run in parallel with the part of an execution plan which precedes the sort box. The second phase uses a multiway merge algorithm to merge all received (and sorted) envelopes into the resulting stream of data lines. In contrary to the first phase, this phase may run in parallel with the part of the execution plan which succeeds the sort box.

4) Merge join: Merge join is a very efficient join algorithm when both inputs are sorted by the common variables. Moreover, the merge join is the algorithm which is suitable for systems like Bobox since it reads both inputs sequentially allowing both input branches to run in parallel (in contrary to hash join, see Section IV-C6).

5) Nested loops join: The SPARQL compiler selects nested loops join when the inputs have no common variable and the result is determined only by the join condition. The implementation is straightforward; however, in order to increase the pipeline parallelism, the box tries to process envelopes immediately as they arrive, i.e. it does not read the whole input before processing the other.

6) Hash join: Hash join is used when the inputs have some common variables which are not sorted in the same order. In order to increase pipeline and task parallelism, we decided not to implement this algorithm. The problem with hash join is that it must read the whole one input first before processing the second one. However, the branch of the plan which produces data for the second input may be blocked because of flow control (see Section II-B) until the first input is completely processed.

Therefore, instead of hash join we implemented sort-merge join. The sort operation is used to transform the inputs to be usable by merge join.

7) Optional joins: Optional join works basically in the same way as regular join. The only difference is that data lines from the left input which do not meet the join condition (i.e., they are not joined with any data line from the right input), are also passed to the output and the variables which come from the right input are set as unbound.

This modification can be easily done when exactly one data line from the left is joined with exactly one data line from the right. In other cases we must keep information about data lines from the left which were already joined and which were not. To do this, each incoming envelope from the left input is extended by one column of boolean values initially set to `false`. When a data line from the left is joined with some data line from the right, we set corresponding boolean value to `true`. When the algorithm finishes, we know which left data lines were not joined and

should be copied to the output.

8) Distinct: Operator distinct should output only unique data lines. We implemented this operator by the modification of a sort operator. The first phase is completely the same; however, during the merging in the second phase, the duplicated data lines are omitted from the output.

9) Other operators: The rest of operators is implemented very straightforwardly. Therefore, we do not describe them here.

D. Explicit parallelization of nested loops join

With the set of boxes described in Section IV-C, we can evaluate the complete SP^{2}Bench benchmark (see Section V). Despite the fact that the implicit parallelization speeds up the evaluation of several queries, this speed up does not scale with the number of physical cores in the host system.

Therefore, we focused on the most time-consuming operation – nested loops join – and tried to explicitly parallelize it using Bobox.

The task of nested loops join is to evaluate the join condition on all pairs of data lines from the left input and data lines form the right input.

This operation can be easily parallelized, since we can create N boxes which perform nested loops join (N denotes the number of worker threads used by Bobox). We pass one N-th of one input and the whole second input to each of these boxes and join their outputs together. It can be easily seen that this modification is valid since all pairs of data lines are still correctly processed. The whole schema of boxes is depicted in Figure 7.

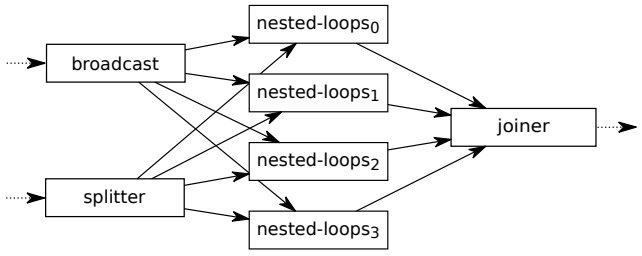


Figure 7. Parallelized nested loops join

The box *splitter* splits its input envelopes to N parts and sends these parts to its outputs. The implementation of this box must be careful since rounding errors may cause that splitted streams do not have the same length. The box *broadcast* just resends its every incoming envelope to its outputs and the box *joiner* resends any incoming envelope to its output.

Since all these three boxes are already implemented in Bobox as standard boxes, the parallelization of nested loops join is very simple.

V. Experiments

We performed a number of experiments to test functionality, performance and scalability of the SPARQL query engine. The experiments were performed using the SP^{2}Bench [16] query set since this benchmark is considered to be a standard in the area of semantic processing.

Experiments were performed on a server running Redhat 6.0 Linux; server configuration is 2x Intel Xeon E5310, 1.60Ghz (L1: 32kB+32kB L2: 4MB shared) and 8GB RAM. It was dedicated specially to the testing; therefore, no other application were running on the server during measurements. SPARQL front-end and Bobox are implemented in C++. Data were stored in-memory.

A. Implicit parallelization

In the first experiment, we measured the speed up caused by the implicit parallelization exploited by Bobox. To measure it, we chose some queries and evaluted them with an increasing number of worker threads. We did not use parallelized version of nested loops join in this experiment and we measured only runtime of evaluation of execution plan, i.e. we did not include the time spent by compilation of the query. The results are shown in Figure 8.

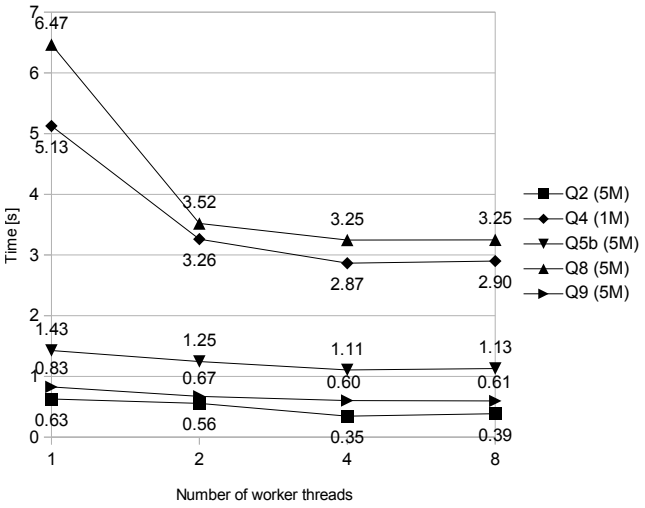


Figure 8. The speed up obtained by implicit parallelization

The results show that for some queries the speed up is quite significant; however, it does not scale with the increasing number of worker threads. This is caused by the fact that the level of parallelism is implicitly built in the execution plan which does not depend on the number of worker threads.

The query Q4 and Q8 benefits from the parallel evaluation most, since the last sort box (or distinct box respectively) runs in parallel with the rest of the execution plan. That is not the case of Q9 which contains distinct box as well; however, the amount of data processed by this box is too small to fully exploit the pipeline parallelism.

	Q1	Q2	Q3a	Q3b	Q3c	Q4	Q5a/b	Q6	Q7	Q8	Q9	Q10	Q11
10k	1	147	846	9	0	23.2k	155	229	0	184	4	166	10
50k	1	965	3.6k	25	0	104.7k	1.1k	1.8k	2	264	4	307	10
250k	1	6.2k	15.9k	127	0	542.8k	6.9k	12.1k	62	332	4	452	10
1M	1	32.8k	52.7k	379	0	2.6M	35.2k	62.8k	292	400	4	572	10
5M	1	248.7k	192.4k	1.3k	0	18.4M	210.7k	417.6k	1.2k	493	4	656	10

Table I
QUERY RESULT SIZES ON DOCUMENTS UP TO 5M TRIPLES.

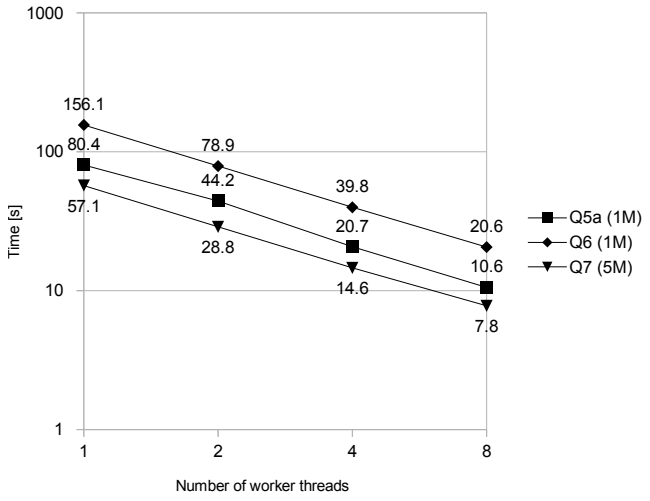


Figure 9. The speed up obtained by explicit parallelization of nested loops join

B. Explicit parallelization

In the second experiment, we focused on the speed up caused by the explicit parallelization of nested loops join. We selected the most time consuming queries with the nested loops joins. As in the first experiment, we performed multiple measurements with the increasing number of worker threads. In this experiment we also did not include the time needed by the query compilation since we focused on the runtime.

The results are shown in Figure 9. According to our expectations, data parallelism increases the scalability and causes a significant almost linear speed up on multiprocessor systems.

C. Comparison with other engines

The last set of experiments compares the Bobox SPARQL engine to other mainstream SPARQL engines, such as Sesame v2.0 [5], Jena v2.7.4 with TDB v0.9.4 [6] and Virtuoso v6.1.6.3127 (multithreaded) [7]. They follow client-server architecture and we provide sum of the times of client and server processes. The Bobox engine was compiled as a single application; we applied timers in the way that document loading times were excluded to be comparable with a server that has data already prepared.

For all scenarios, we carried out multiple runs over documents containing 10k, 50k, 250k, 1M, and 5M triples and we provide the average times. Each test run was also limited to 30 minutes (the same timeout as in the original SP^{2}Bench paper). All data were stored in-memory, as our primary interest is to compare the basic performance of the approaches rather than caching etc. The expected number of the results for each scenario can be found in Table I.

The query execution times are shown in Figure 10. The y-axes are shown in a logarithmic scale and individual plots scale differently. In the following paragraphs, we discuss some of the queries and their results. In contrary to previous experiments, we did include the time spent by the compiler in order to be comparable with other engines.

Q2 implements a bushy graph pattern and the size of the result grows with the size of the queried data. We can see that Bobox Engine scales well, even though it creates execution plans shaped as a left-deep tree. This is due to the parallel stream processing of merge joins. The reason why our solution is slower on 10k and 50k of triples is that the compiler takes more than 1s to compile and to optimize the query.

The variants of Q3 (labelled *a* to *c*) test FILTER expression with varying selectivity. We present only the results of Q3c as the results for Q3a and Q3b are similar. The performance of Bobox is negatively affected by a simple implementation of statistics used to estimate the selectivity of the filter.

Q4 (Figure 11) contains a comparably long graph chain, i.e., variables `?name1` and `?name2` are linked through articles that (different) authors have published in the same journal. Bobox embeds the FILTER expression into this computation instead of evaluating the outer pattern block and applying the FILTER afterwards and propagates the DISTINCT modifier closer to the leaves of the plan in order to reduce the size of the intermediate results.

Queries Q5 (Figure 11) test implicit (Q5a) join encoded in a FILTER condition and explicit (Q5b) variant of joins. On explicit join both engines used fast join algorithm and are able to produce result in a reasonable time. On implicit join both engines used nested loops join which scales very badly. However, Bobox outperforms both Sesame and Jena since it is able to use multiple processors to get the results and is able to compute also documents with 250k, 1M and 5MB

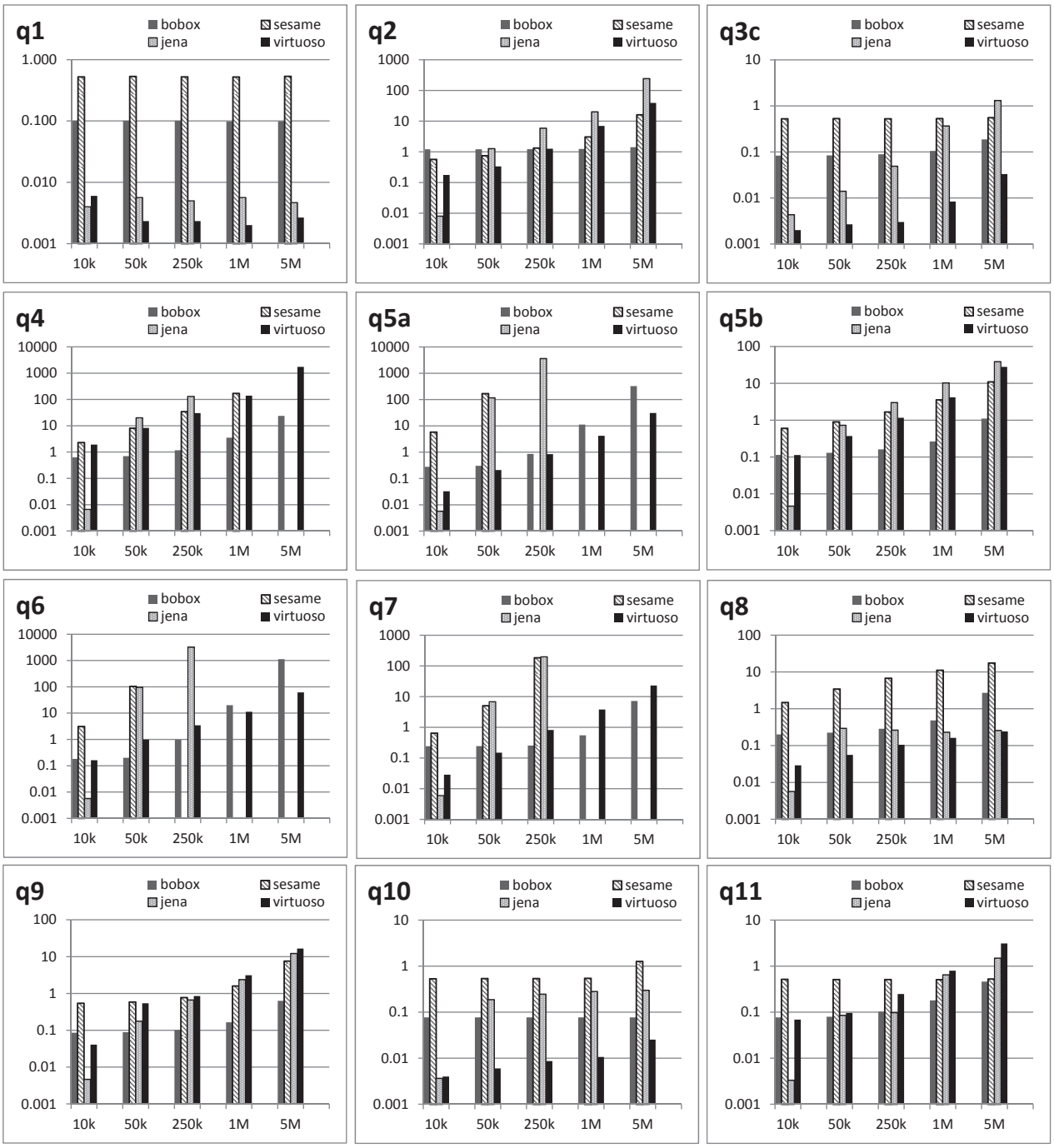


Figure 10. Results (time in seconds) for 10k, 50k, 250k, 1M, and 5M triples.

triples before the 30 minute limit is reached. On the other hand, Virtuoso outpeforms Bobox mainly due to particular query optimizations [16].

Queries Q6, Q7 and Q8 produce bushy trees; their computation is well handled in parallel, mainly because of nested loops join parallelization. As a result of this, Bobox outperforms Sesame and Jena in Q6 and Q7 and outperforms Virtuoso in Q7, being able to compute larger documents until the query times out. The authors of the SP^{2}Bench suggest [16] reusing graph patterns in a description of the queries Q6, Q7 and Q8. However, this is problematical in Bobox. Bobox processing is driven by the availability of the data on inputs but it also incorporates methods to prevent the input buffers from being overfilled (see Section II-B).

```
SELECT DISTINCT ?name1 ?name2                                    Q4
WHERE { ?article1 rdf:type bench:Article.
        ?article2 rdf:type bench:Article.
        ?article1 dc:creator ?author1.
        ?author1 foaf:name ?name1.
        ?article2 dc:creator ?author2.
        ?author2 foaf:name ?name2.
        ?article1 swrc:journal ?journal.
        ?article2 swrc:journal ?journal
        FILTER (?name1<?name2) }

SELECT DISTINCT ?person ?name                                    Q5a
WHERE { ?article rdf:type bench:Article.
        ?article dc:creator ?person.
        ?inproc rdf:type bench:Inproceedings.
        ?inproc dc:creator ?person2.
        ?person foaf:name ?name.
        ?person2 foaf:name ?name2
        FILTER(?name=?name2) }
```

Figure 11. Examples of the benchmark queries.

Pattern reusing can result in the same data being sent along two different paths in the pipeline running at a different speed. Such paths may then converge in a join operation. When the faster path overfills the input buffer of the join box, the computation of all boxes on paths leading to the box is suspended. As a result, data for the slower path will never be produced and will not reach the join box, which results in a deadlock. We intend to examine the possibility of introducing a buffer box, which will be able to store and provide data on request. This way, the Bobox SPARQL implementation will be able to reuse graph patterns.

Q10 can be processed very fast because of our database representation. Therefore, only resulting triples are fetched directly from the database.

In contrary to Sesame, time of Q11 depends on the size of database. This is caused by the fact that we do not have any optimization for queries with LIMIT or OFFSET modifiers. In that case, the whole results set is produced which naturally slows down the evaluation.

Overall, the results of the benchmarks indicate very good potential of Bobox when used for implementation of RDF query engine. Our solution outperforms in all measurements Sesame, in most cases significantly and in most measurements Jena. The performance of Bobox and Virtuoso is comparable; Bobox outperforms Virtuoso namely in computing and data intensive queries.

VI. Related work

A. Parallel Frameworks and Libraries

The most similar to the Bobox run-time is the TBB library [26]. It was one of the first libraries that focused on task level parallelism. Compared to the Bobox, it is a low-level solution – it provides basic algorithms like parallel for cycle or linear pipeline and a very efficient task scheduler. The developers are able to directly create tasks for the scheduler and create their own parallel algorithms. But the tasks are designed in a way that makes it very hard to create a non-linear pipeline similar to the one Bobox provides. Such pipeline may be necessary for complex data processing [27]. Bobox also provides more services for data passing and flow control.

The latest version of OpenMP [28] also provides a way to execute tasks in parallel, but it provides less features and less control than TBB. The OpenMP library is mainly focused on mathematical computations – it can execute simple loops in parallel really fast, it can also run blocks of code in parallel, but it is not well suited for parallel execution of a complex structure of blocks. Unlike TBB or Bobox, it is a language extension and not just a library; the compiler is well aware of the parallelization and optimize the code better, but it also enables OpenMP to provide features that cannot be done with just a library, like defining the way variables are shared among threads with a simple declaration. In TBB such variable has to be explicitly passed to an appropriate algorithm by the programmer. In Bobox, it must either be explicitly passed to the model or sent using an envelope at run-time.

Some of the architectural decisions could be implemented in a different manner. One way would be to create a thread for each box and via in the model instance. This would also ensure that each box or via is running at most once at any given time. However, this is considered a bad practice [29]. There are two main reasons for not using this architecture. First, it creates a large number of threads, usually much larger than the number of CPU cores. Although it forces the operating system to switch the threads running on a core, it may not impact the overall performance that badly, since it can be arranged that the idling threads (those assigned to a box or via that is not processing any data at the moment) are suspended and do not consume any CPU time. The second problem is that when data (envelopes) are transfered from one box to another, there is very little chance that it would still be hot in the cache, since the thread that corresponds to the second box is likely to be scheduled to a different CPU, that does not share its cache with the original one. The concept of tasks used by TBB and Bobox avoid these problems and the use of thread pool, fixed number of threads and explicit scheduling gives developers of the libraries better control of parallel execution.

Besides these low-level techniques of parallel data processing, the *MapReduce* approach gained significant attention. While it is often considered a step back [30], there are application areas where MapReduce may outperform a parallel database [31]. Although MapReduce was originally targeted to other environments, it was also studied in shared-memory settings (similar to Bobox) [32], [33]. Unlike MapReduce, Bobox is desiged to support more complicated processing environment, namely nonlinear pipelines.

B. Parallel Databases

In a relational database management system, parallelism may be employed at various levels of its architecture:

- *Inter-transaction parallelism.* Running different transactions in parallel has been a standard practice for decades. Besides dealing with disk latency, it is also the easiest way to achieve a degree of parallelism in shared-memory or shared-disk environment. Although it is not considered a specific feature of parallel databases, it must be carefully considered in the design of parallel databases since parallel transactions compete for memory, cache, and bandwidth resources [34], [35].
- *Intra-transaction parallelism.* Queries of a transaction may be executed in parallel, provided they do not interfere among themselves and they do not interact with external world. Since these conditions are met rather rarely, this kind of parallelism is seldom exploited except for experiments [36].
- *Inter-operator parallelism.* Since individual operators of a physical query plan have well-defined interfaces and mostly independent behavior, they may be arranged to run in parallel relatively easily. On the other hand, the effect of such parallelism is limited because most of the cost of a query plan is often concentrated in one or a few of the operators [37].
- *Intra-operator parallelism.* Parallelizing the operation of a single physical operator is the central idea of parallel databases. From the architectural point of view, there are two different approaches:
  - a) *Partitioning* [38] – this technique essentially distributes the workload using the fact that many physical algebra operators are distributive with respect to union (or may be rewritten using such operators).
  - b) *Parallel algorithms* – implementing the operator using a parallel algorithm usually offers the freedom of control over the time and resource sharing and machine-specific means like atomic operations or SIMD instructions. However, designing, implementing, and tuning a parallel algorithm is an extremely complex task, often producing errors or varying performance results [39]. Moreover, the evolution of hardware may soon make a parallel implementation obsolete [40]. For these reasons, parallelizing frameworks are developed [41].

The central principle of Bobox allows parallelism among boxes but prohibits (thread-based) parallelism inside a box. This is similar to inter-transaction and inter-operator parallelism; however, a box does not necessarily correspond to a relational operator. In particular, Bobox allows the same approach to partitioning as in parallel databases, using transformation of the query plan.

Bobox does not allow parallel algorithms to be implemented inside a box (except of the use of SIMD instructions). Therefore, individual single-threaded parts of a parallel algorithm must be enclosed in their boxes and the complete algorithm must be built as a network of these boxes. This is certainly a limitation in the expressive power of the system; on the other hand, the communication and synchronization tasks are handled automatically by the Bobox framework.

VII. Conclusions and Future Work

In the paper, we presented a parallel SPARQL processing engine that was built using the Bobox framework with a focus on efficient query processing: parsing, optimization, transformation and parallel execution. We also presented the parallelization of nested loops join algorithm to increase parallelism during the evaluation of time consuming queries. Despite the fact that this parallelization is very simple to be done using Bobox, the measurements show that it scales very well in a multiprocessor environment.

To test the performance, we performed multiple sets of experiments. We have chosen established frameworks for RDF data processing as the reference systems. The results seem very promising; using SP^{2}Bench queries we have identified that our solution is able to process many queries significantly faster than other engines and to obtain results on larger datasets. Therefore, such a parallel approach to RDF data processing has a potential to provide better performance than current engines. On the other hand, we also identified several issues:

- We are working on improvements of our statistics used by the compiler to generate more optimal query plans.
- The pilot implementation of the compiler is not well optimized which is problem especially in Q1 and Q2.
- Our heuristics sometimes result in long chains of boxes. Streamed processing and fast merge joins minimize this disadvantage; however, it is better to have bushy query plans for efficient parallel evaluation.
- Also, some methods proposed in SP^{2}Bench, such as graph pattern reuse, are not efficiently applicable in the current Bobox version.
- The query Q4 is very time consuming and does not benefit much from the fact that the system has multiple processors. Therefore, we must parallelize besides the nested loops join also merge join, which is the bottleneck of this query.
- Currently, we support only in-memory databases. In order to have engine usable for processing of really large RDF databases such as BTC Dataset (Billion triple challenge) [42], we must keep the database in external memory.

Because of these issues, we are convinced that there is still space for optimization in parallel RDF processing and we want to focus on them and improve our solution.

ACKNOWLEDGMENTS

The authors would like to thank the GAUK project no. 28910, 277911 and SVV-2012-265312, and GACR project no. 202/10/0761, which supported this paper.

REFERENCES

[1] M. Cermak, J. Dokulil, Z. Falt, and F. Zavoral, "SPARQL Query Processing Using Bobox Framework," in *SEMAPRO 2011, The Fifth International Conference on Advances in Semantic Processing*. IARIA, 2011, pp. 104–109.

[2] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF," W3C Recommendation, 2008.

[3] J. J. Carroll and G. Klyne, *Resource Description Framework: Concepts and Abstract Syntax*, W3C, 2004. [Online]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

[4] Y. Yan, C. Wang, A. Zhou, W. Qian, L. Ma, and Y. Pan, "Efficiently querying rdf data in triple stores," in *Proceeding of the 17th international conference on World Wide Web*, ser. WWW '08. New York, NY, USA: ACM, 2008, pp. 1053–1054.

[5] J. Broekstra, A. Kampman, and F. v. Harmelen, "Sesame: A generic architecture for storing and querying RDF and RDF schema," in *ISWC '02: Proceedings of the First International Semantic Web Conference on The Semantic Web*. London, UK: Springer-Verlag, 2002, pp. 54–68.

[6] "Jena – a semantic web framework for Java," http://jena.sourceforge.net. [Online]. Available: http://jena.sourceforge.net, retrieved 10/2012

[7] "Virtuoso data server," http://virtuoso.openlinksw.com, retrieved 10/2012

[8] A. Kiryakov, D. Ognyanov, and D. Manov, "Owlim a pragmatic semantic repository for owl," 2005, pp. 182–192.

[9] T. Neumann and G. Weikum, "The rdf-3x engine for scalable management of rdf data," *The VLDB Journal*, vol. 19, pp. 91–113, February 2010.

[10] J. Huang, D. Abadi, and K. Ren, "Scalable sparql querying of large rdf graphs," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, 2011.

[11] Z. Falt, D. Bednarek, M. Cermak, and F. Zavoral, "On Parallel Evaluation of SPARQL Queries," in *DBKDA 2012, The Fourth International Conference on Advances in Databases, Knowledge, and Data Applications*. IARIA, 2012, pp. 97–102.

[12] D. Bednarek, J. Dokulil, J. Yaghob, and F. Zavoral, "Data-Flow Awareness in Parallel Data Processing," in *6th International Symposium on Intelligent Distributed Computing - IDC 2012*. Springer-Verlag, 2012.

[13] "The Bobox Project - Parallelization Framework and Server for Data Processing," 2011, Technical Report 2011/1. [Online]. Available: http://www.ksi.mff.cuni.cz/bobox, retrieved 12/2012

[14] M. Schmidt, T. Hornung, N. Küchlin, G. Lausen, and C. Pinkel, "An Experimental Comparison of RDF Data Management Approaches in a SPARQL Benchmark Scenario," in *ISWC*, Karlsruhe, 2008, pp. 82–97.

[15] D. Bednarek, J. Dokulil, J. Yaghob, and F. Zavoral, "Bobox: Parallelization Framework for Data Processing," in *Advances in Information Technology and Applied Computing*, 2012.

[16] M. Schmidt, T. Hornung, G. Lausen, and C. Pinkel, "Sp2bench: A sparql performance benchmark," *CoRR*, vol. abs/0806.4627, 2008.

[17] Z. Falt, J. Bulanek, and J. Yaghob, "On Parallel Sorting of Data Streams," in *ADBIS 2012 - 16th East European Conference in Advances in Databases and Information Systems*, 2012.

[18] J. Dokulil and J. Katreniakova, "Bobox model visualization," in *14th International Conference Information Visualisation*. London, UK: IEEE Computer Society, 2010, pp. 537–542.

[19] D. Bednarek, J. Dokulil, J. Yaghob, and F. Zavoral, "Using Methods of Parallel Semi-structured Data Processing for Semantic Web," in *3rd International Conference on Advances in Semantic Processing, SEMAPRO*. IEEE Computer Society Press, 2009, pp. 44–49.

[20] J. Galgonek, "Tequila - a query language for the semantic web," in *DATESO 2009*, ser. CEUR Workshop Proceedings, K. Richta, J. Pokorný, and V. Snášel, Eds., vol. 471. Czech Technical University in Prague, 2009, pp. 105–118.

[21] M. Krulis and J. Yaghob, "Revision of relational joins for multi-core and many-core architectures," in *Proceedings of the Dateso 2011*. Pisek, Czech Rep.: FEECS, 2011.

[22] Z. Falt and J. Yaghob, "Task Scheduling in Data Stream Processing," in *Proceedings of the Dateso 2011 Workshop*. Citeseer, 2011, pp. 85–96.

[23] H. Pirahesh, J. M. Hellerstein, and W. Hasan, "Extensible/rule based query rewrite optimization in starburst," *SIGMOD Rec.*, vol. 21, pp. 39–48, June 1992.

[24] O. Hartig and R. Heese, "The SPARQL Query Graph Model for query optimization," in *The Semantic Web: Research and Applications*, ser. Lecture Notes in Computer Science, E. Franconi, M. Kifer, and W. May, Eds. Springer Berlin / Heidelberg, 2007, vol. 4519, pp. 564–578.

[25] M. Cermak, J. Dokulil, and F. Zavoral, "SPARQL Compiler for Bobox," *Fourth International Conference on Advances in Semantic Processing*, pp. 100–105, 2010.

[26] A. Kukanov and M. J. Voss, "The foundations for scalable multi-core software in Intel Threading Building Blocks," *Intel Technology Journal*, vol. 11, no. 04, pp. 309–322, November 2007.

[27] D. Bednárek, "Bulk evaluation of user-defined functions in XQuery," Ph.D. dissertation, Department of Software Engineering, Faculty of Mathematics and Physics, Charles University in Prague, 2009.

[28] *OpenMP Application Program Interface, Version 3.0*, OpenMP Architecture Review Board, May 2008, http://www.openmp.org/mp-documents/spec30.pdf, retrieved 9/2011.

[29] J. Reinders, *Intel Threading Building Blocks*. O'Reilly, 2007.

[30] M. Stonebraker, D. Abadi, D. J. DeWitt, S. Madden, E. Paulson, A. Pavlo, and A. Rasin, "Mapreduce and parallel dbmss: friends or foes?" *Commun. ACM*, vol. 53, pp. 64–71, 2010.

[31] S. Blanas, J. M. Patel, V. Ercegovac, J. Rao, E. J. Shekita, and Y. Tian, "A comparison of join algorithms for log processing in mapreduce," in *SIGMOD '10: Proceedings of the 2010 international conference on Management of data*. USA: ACM, 2010, pp. 975–986.

[32] C. Ranger, R. Raghuraman, A. Penmetsa, G. Bradski, and C. Kozyrakis, "Evaluating mapreduce for multi-core and multiprocessor systems," in *HPCA '07: Proceedings of the 2007 IEEE 13th International Symposium on High Performance Computer Architecture*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 13–24.

[33] G. Kovoor, "MR-J: A MapReduce framework for multi-core architectures," Ph.D. dissertation, University of Manchester, 2009.

[34] F. Morvan and A. Hameurlain, "Dynamic memory allocation strategies for parallel query execution," in *SAC '02: Proceedings of the 2002 ACM symposium on Applied computing*. New York, NY, USA: ACM, 2002, pp. 897–901.

[35] Z. Zhang, P. Trancoso, and J. Torrellas, "Memory system performance of a database in a shared-memory multiprocessor," 2007. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.1924, retrieved 10/2012

[36] C. B. Colohan, A. Ailamaki, J. G. Steffan, and T. C. Mowry, "Optimistic intra-transaction parallelism on chip multiprocessors," in *VLDB '05: Proceedings of the 31st international conference on Very large data bases*. VLDB Endowment, 2005, pp. 73–84.

[37] A. N. Wilschut, J. Flokstra, and P. M. G. Apers, "Parallel evaluation of multi-join queries," in *SIGMOD '95: Proceedings of the 1995 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM, 1995, pp. 115–126.

[38] D. DeWitt and J. Gray, "Parallel database systems: the future of high performance database systems," *Commun. ACM*, vol. 35, no. 6, pp. 85–98, 1992.

[39] J. Aguilar-Saborit, V. Muntes-Mulero, C. Zuzarte, A. Zubiri, and J.-L. Larriba-Pey, "Dynamic out of core join processing in symmetric multiprocessors," in *PDP '06: Proceedings of the 14th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 28—35.

[40] C. Kim, T. Kaldewey, V. W. Lee, E. Sedlar, A. D. Nguyen, N. Satish, J. Chhugani, A. Di Blas, and P. Dubey, "Sort vs. hash revisited: fast join implementation on modern multi-core cpus," *Proc. VLDB Endow.*, vol. 2, no. 2, pp. 1378–1389, 2009.

[41] J. Cieslewicz, K. A. Ross, K. Satsumi, and Y. Ye, "Automatic contention detection and amelioration for data-intensive operations," in *SIGMOD '10: Proceedings of the 2010 international conference on Management of data*. New York, NY, USA: ACM, 2010, pp. 483–494.

[42] "Billion triple challenge." [Online]. Available: http://challenge.semanticweb.org, retrieved 10/2012

Fuzzy Query Propagation in Sensor Networks

Mohamed Bakillah Steve H.L. Liang
Department of Geomatics Engineering
University of Calgary, 2500 University Dr. NW, Canada
mohamed.bakillah@geog.uni-heidelberg.de
steve.liang@ucalgary.ca

Mir Abolfazl Mostafavi
Geomatic Research Center, Laval University,
1055, avenue du Séminaire, Québec,
Canada
mir-abolfazl.mostafavi@scg.ulaval.ca

Alexander Zipf Jamal Jokar Arsanjani
Rupprecht-Karls-Universität, Institute for GI-Science,
Berliner Straße 48, D-69120, Heidelberg, Germany
alexander.zipf@geog.uni-heidelberg.de
jokar.arsanjani@geog.uni-heidelberg.de

***Abstract*—Query-driven information retrieval aims at supporting users to request and retrieve relevant data from sensor networks. Due to energy and capacity constraints that characterize sensor networks, information retrieval should avoid flooding the network with queries, but rather find the most efficient propagation path that maximizes the recall of relevant data while minimizing the number of sensor nodes being accessed. This is the problem of query propagation, for which numerous approaches for sensor networks have been proposed. Although, one unaddressed issue that remains is the issue of fuzziness of users' queries and fuzziness of sensor data. When crisp criteria are used to express queries and select query recipients during propagation, some sensor nodes that are relevant can be "missed." Therefore, this paper's objective is to integrate a fuzzy semantic mapping mechanism, which has been published in a previous research, into a new, cluster-based fuzzy query propagation approach. The fuzzy query propagation approach avoids the overload of sensor nodes that are near the sink nodes by incorporating a first propagation step towards relevant clusters of sensors, therefore varying the sensor nodes that will have to redistribute the query, followed by an intra-cluster query propagation phase. The approach has been evaluated with a simulation and compared with a crisp version to show the impact of the consideration of fuzziness in the improvement of the recall of relevant nodes while avoiding the increase of propagation cost.**

Keywords - Fuzzy logics, information retrieval, query propagation, sensor networks.

I. INTRODUCTION

Sensor networks are intended to monitor environmental conditions, such as weather, properties of soil and water bodies, vegetation, etc. While sensors are more traditionally used by scientific experts to study environmental and physical phenomena, it is believed that greater public can also benefit from access to sensor data. In this paper, we extend a previous paper on fuzzy semantic mapping that was presented at the SEMAPRO 2011 conference [1] by integrating the fuzzy semantic mapping approach into a new fuzzy query propagation approach for sensor networks.

In order to support improved access, sensor data should therefore be accessed through the Internet, with the help of platforms such as the Geospatial Cyberinfrastructure for Environmental Sensing platform (GeoCENS). GeoCENS is an online platform that enables simplified searching, storing and sharing of environmental and other georeferenced data [2]. In such platform, sensors collect data on a given feature, process these data and forward it to a so-called "sink node," which in turn forwards the data to the application through the Internet. Because all sensor nodes cannot necessarily be connected to the sink node, sensor data must be forwarded from node to node until reaching the sink node [3]. In the same manner, sensor data queries issued by users must be forwarded from the sink node to the nodes holding the requested data (the relevant nodes) through intermediary sensor nodes in the network.

However, because sensors are meant to be small devices, their processing capacity and their source of energy are limited. Also, despite the decreasing cost of sensors, it cannot be assumed that they can be replaced when they run out of power. For example, some sensors cannot be accessed once being set up in their environment (some are buried to measure soil moisture, while others are underwater to measure water temperature, etc.). Therefore, the path chosen to send queries to sensor nodes and to send back data to the sink node must be determined in a way to avoid consuming the energy of sensor nodes; at the same time, the path chosen must enable to reach the nodes that are relevant to the query and retrieve the requested data. This problem is called query propagation.

Numerous approaches have been proposed for query propagation and data collection from sensor nodes. A representative sample of such approach is presented in Section II. The approaches are varying in terms of the data delivery model (whether sensors proactively send data to the sink node according to a pre-defined scheme, or solely on-demand of the user); organization of the sensor network (flat

or hierarchical); and criteria for selecting query recipient nodes. However, one well-know problem in GIScience, but that is still unaddressed in query propagation approaches for sensor networks, is the fuzziness of data and queries.

Research indicates that geographical phenomena in particular are fuzzy [4]. For example, where a mountain starts or ends cannot be determined with precision; and whether the vegetation is dense or not is only an imprecise concept. Fuzzy theory, which allows the partial membership of an element into a set (e.g., the set of dense vegetation areas) is widely used to represent geographical phenomena. For example, in [5], fuzzy theory is used to represent fuzzy land cover categories. Similarly, concepts such as spatial relations that are used in users' queries (e.g., close to, around, at proximity, far from) are also fuzzy [6]. Sensor data can also be fuzzy, for example, the location of the sensor can be imprecise or there is a certain level of uncertainty in data being gathered. The fact that sensor data and queries are fuzzy should be taken into account during query propagation. Conversely, it could result in the inability of the approach to retrieve relevant data.

In previous research presented at the SEMAPRO 2011 conference [1], we have presented a fuzzy logic semantic mapping model to compare components of fuzzy ontologies. In this paper, the objective is to apply this approach and integrate it into a fuzzy query propagation approach. The fuzzy semantic mapping theory and mechanism presented in [1] is incorporated into a cluster-based query propagation approach as a way to express fuzzy queries and select query recipient according to fuzziness degree and a fuzzy semantic relations. The fuzzy query propagation approach incorporates a first propagation step towards relevant clusters of sensors, followed by an intra-cluster query propagation phase. The ability of the approach to retrieve relevant information and a comparison between crisp and fuzzy propagation has been evaluated through a simulation.

The content of this paper is organized as follows: the next section presents related work on query propagation in sensor networks. Section III is a brief introduction to fuzzy logics in GIScience and in sensor networks. Section IV presents our fuzzy query propagation approach, while Section V presents an extended version of the fuzzy semantic mapping mechanism. The evaluation of the approach is conducted in Section VI, while conclusion and future work are provided in Section VII.

II. QUERY PROPAGATION IN SENSOR NETWORKS

Propagating queries to the relevant sensors of a network is a challenging issue, since a balance between the quality of query answers and the efficiency of the approach must be reached.

Existing query propagation approaches for sensor networks can be categorized according to the data delivery model they rely on, i.e., how the flow of data between the sensors and the requestor is triggered and organized [7]. The first data delivery model is the proactive model. In the proactive model, sensor nodes periodically forward the data they have collected to a server, at a pre-specified rate, or when an event of interest occurs (event-driven model) [3]. Examples of query propagation approaches based on the proactive data delivery model include [8] and [9]. While the approach proposed in this paper could be somewhat easily adapted to the proactive model, in this paper, we focus on the second type of delivery model, i.e., the query-driven model.

In the query-driven model (or on-demand model), data is sent by sensor nodes only when a user queries the sensor network [10]. The problem then is to determine through which path and to which sensor nodes the query should be sent. We assume in the following that the user can access the sensor network through a so-called “sink node,” which is a node of the network that acts as an intermediary between the user (through the Internet) and the rest of sensors in the network [11]. One common approach for query-driven model is the reverse tree model [10][12][13]. In the reverse tree model, the query is broadcasted from the sink node to the nodes of the network. The structure of the tree is built as the query is propagated from node to node, with the sink node being the root of the tree. Sensors send back their data to the sink node following the tree structure. Approaches based on the reverse tree model vary according to the mechanism they rely on for selecting the nodes that will be part of the tree. For example, some approaches are called “attribute-based,” because at each “jump,” the decision about propagation is made based on a match between the attributes specified in the query and the attributes of data collected by the sensors. Examples of such approaches include [14][15][16][10]. The attribute can be, for example, the area where the sensor is located or the type of sensor. One disadvantage of the reverse tree model is that it can be inefficient because it may impose unbalanced energy consumption in the sensor network, since the nodes that are close to the sink forward more data and queries and therefore, use more energy than other nodes that are far from the sink node [10]. One solution would be therefore to avoid that the sink node always sends the query through its immediate neighbors. To address this issue, and to facilitate routing to relevant sensors in general, the hierarchical routing protocols can be helpful. Hierarchical routing protocols divide the network into clusters of sensors [17][18][8][9]. Queries can then be sent directly from the sink node to the designated "leader" of the relevant cluster, avoiding the same sensors to disseminate the queries and collect the corresponding data.

Other types of approaches, called geographical routing protocols, aim at propagating the queries sent by users who are searching for data from sensors in a specific location. These protocols therefore explicitly take into account the location of sensors in the selection of recipient nodes [19][20][21]. The query includes the targeted coordinates; neighbor sensor nodes in the network are actively sharing the information about their respective location. Therefore, when a node receives a query, it sends it to the neighbor node that is the closest to the targeted location. Villalba et al. [3] indicate that several metrics have been used to measure closeness, the most common ones being the Euclidean distance and the projected line joining the relaying node and the destination. However, we note that such routing protocols based on crisp measures do not allow take into

account the fuzziness of queries. More particularly, it is very likely that users lack the capacity to specify a precise location of interest, and can only provide an approximation of it [10]. We argue that this is also true regarding thematic or temporal attributes of queries. For example, a user might look for sensors that have observed "temperature around 30°C" rather than exactly 30°C, within a fuzzy period of time. This motivates our proposal of a fuzzy query propagation approach for sensor networks.

III. FUZZY LOGICS IN GISCIENCE AND SENSOR NETWORKS

GIScience researchers such as Couclecis [22] and Zhang and Goodchild [23] have demonstrated that uncertainty should be considered as a kind of knowledge that must be explicitly represented and dealt with. Fuzzy logics, which were proposed by Zadeh [24] to deal with imprecise and vague knowledge, are now widely used in GIScience [4]. For example, [25] uses fuzzy sets to assess the similarity of categorical maps, while [6] have developed an ontology of fuzzy spatial relations to improve the interpretation of images. Fuzzy theory and fuzzy logics are also widely used in sensor networks. For example, [26] use fuzzy logics in a hierarchical clustering protocol for query routing in sensor networks. In this approach, fuzzy logics are used to select the sensor node that will play the role of the "cluster head" (leader) of a sensor cluster. Fuzzy variables used for cluster head selection include energy, centrality, and concentration. Fuzzy logics have also been used to assess the quality of service (QoS) in wireless sensor networks [27]. More specifically, QoS in wireless sensor networks is highly related to energy efficiency and avoiding the congestion of messages at nodes. In [27], fuzzy logics are used to estimate the congestion at nodes in order to facilitate routing messages more efficiently. Fuzzy logics are also used to assess trust in order to distinguish between trustworthy and threatening nodes in wireless sensor networks [28]. In [29], fuzzy theory is used to enable fusion of uncertain sensor data in wireless sensor networks. The approach was designed for the fusion of data coming from sensors that monitor the same property (in this case, luminosity). Other applications exist that use fuzzy theory in the context of message routing in wireless sensor networks [30][31]. [30] propose a solution to avoid the useless propagation of messages to all nodes of the network. In their approach, the transmission area is limited according to a fuzzy threshold value. The fuzzy threshold value is determined by a fuzzy rule-based system that considers the energy and density of nodes. [31] have developed a fuzzy logic controller that allows nodes in the sensor network to compute their capacity to transfer messages based on their battery power level and the type of data being forwarded. Similarly, [32] proposed an energy-aware fuzzy routing mechanism for wireless sensor networks. Despite numerous works using fuzzy theory for sensor networks, to the best of our knowledge, none investigate the use of fuzzy logics to represent the uncertainty of semantics of sensor data and to support semantic-based query propagation. This motivates the approach presented in this paper.

A. Fundamentals of Fuzzy Theory

This section briefly introduces the basic notions of fuzzy sets and fuzzy logics. In classical set theory, elements of a set either belong to a set, or they do not; conversely, fuzzy set theory was developed to deal with the case of partial membership to a set. Each member of a fuzzy set is assigned a so-called membership degree, which value is between 0 and 1, and which indicates the strength of the membership into the set. A null value indicates that the element does not belong at all to the set, while a value of 1 indicates that the element fully belongs to the set. Consider a set of elements called the reference set and denoted X. A fuzzy subset F of X is formally defined with a membership function $\mu_F(x)$; this function associates any element x of X to a value in the [0, 1] interval. All set operations for crisp sets (union, intersection, etc.) have their fuzzy counterpart. The fuzzy implication operators such as Gödel, Gogen and Lukasiewicz fuzzy implications operators are for example used to reason with relations between fuzzy sets [33] while fuzzy composition operators are used to infer membership of an element into a fuzzy set, knowing its membership degree into another related fuzzy set. The operators that will be used in this paper will be introduced in Section V.

IV. FUZZY QUERY PROPAGATION

The data delivery model targeted by the proposed fuzzy query propagation approach is query-driven [3], i.e., the fuzzy query propagation process is initiated by a user who issues a query expressing the characteristics of the data he or she is looking for. Figure 1 illustrates the fuzzy query propagation framework.

The proposed framework is based on the principle of hierarchical routing protocol [3], which advantage is to avoid large traffic overhead and therefore to reduce energy consumption by sensors [10]. In this paper, we assume that the sensor network is already partitioned into clusters of sensors. Each cluster has a gateway node, which is the node responsible for receiving a query and redistributing it to other members of the cluster. Existing research [26] demonstrates that a single gateway node has disadvantages because it can become a single point of failure (e.g., if the selected gateway node runs out of power or becomes dysfunctional). To avoid this problem, the role of gateway node is rotated among several nodes (provided that they have sufficient capacity). The choice of gateway nodes can be done randomly at predetermined time intervals [34] in order to share the consumption of energy. However, in case of failure, the sink node should automatically forward the queries to the next gateway node. To detect failure of the gateway node, we have included a communication protocol where the gateway node sends a notification to the sink node every time it receives a query. Therefore, if the sink node does not receive a notification, it assumes that the current

gateway node is not available and rotate to the next available gateway node.

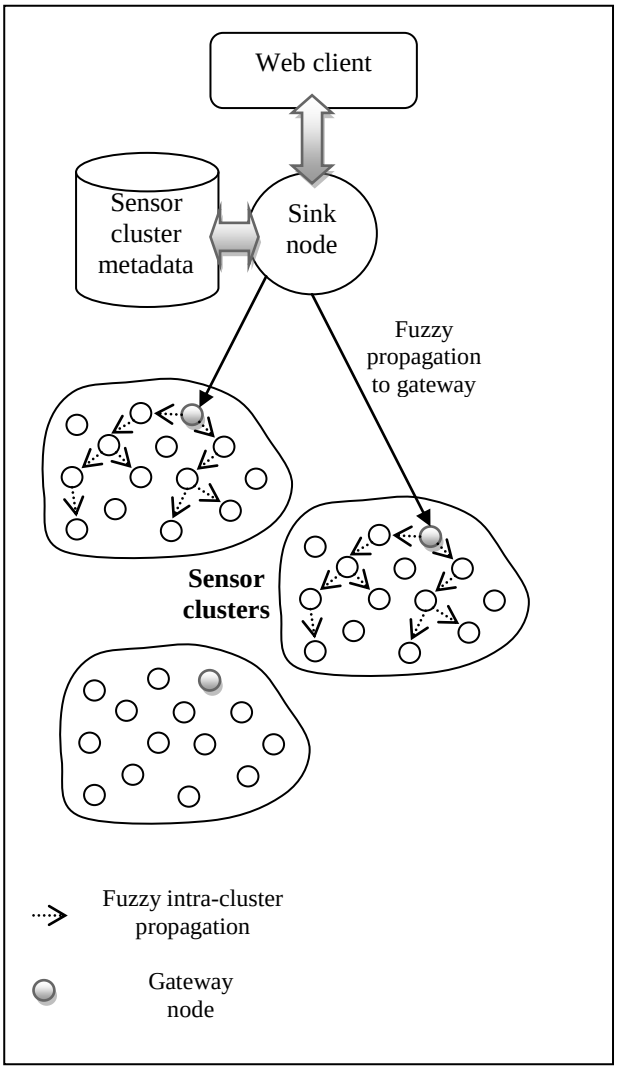


Figure 1. Fuzzy propagation framework

Sensor clusters are formed according to various semantic criteria: sensors which data pertain to similar or complementary domain of application, themes, geographical locations, etc., are gathered into clusters. This facilitates the propagation of queries to targeted groups of sensors instead of flooding the network with queries. Since it is not the objective of this paper to further describe how clusters of sensors are formed, we point out to our previous published research [35] where we have proposed a social-network-analysis-based algorithm for sensor cluster formation. The clustering algorithm identifies, within the available sensors, those that can be considered as "leaders" because their characteristics encompass those of other sensors. For example, a sensor that "measures density of gas" encompasses sensors that "measure density of CO2", "measure density of air pollutant," etc. Leader sensors are identified using the network analysis concept of "centrality." Then, meaningful clusters of sensors are formed around those "leader" sensors. To do so, the algorithm searches the semantic neighborhood of leader sensors to select those that will be part of the cluster formed around this leader sensor.

Each sensor node stores a set of metadata according to the Sensor Model Language (SensorML) format [36]. Similarly, each sensor cluster is associated with metadata that describe the nature of its member sensors. The metadata describe the category of phenomenon observed, the observation period and area of observations, the observed properties (e.g., temperature, soil moisture, etc.), the types of sensors, the intended application and the application domain [36]. The metadata on sensor clusters are stored in a sensor cluster metadata knowledge base, which is held by the sink node that usually has greater storage and processing capabilities than "regular" sensor nodes [10].

A. Global Fuzzy Query Propagation Process

The fuzzy query propagation process is as follows: first, a fuzzy query is formulated by a user. The query is sent, through a Web platform, to a sink node. The sink node is responsible for broadcasting the query to sensors of the network. However, instead of flooding the network, the sink node identifies the clusters that are the most likely to contain sensor nodes that are relevant to the query. To do so, a fuzzy semantic matcher (described in Section V) is implemented at the sink node. The fuzzy semantic matcher compares the fuzzy query with the metadata on the cluster and return matches. Matches are selected according to fuzzy criteria, which computation is discussed in Section V. When matching clusters are selected, the sink node then sends the query to the gateway node of these clusters. Then, the gateway node will initiate the fuzzy intra-cluster propagation of the query, i.e., propagation from node to node inside a cluster.

B. Fuzzy Intra-Cluster Propagation Algorithm

The fuzzy intra-cluster propagation algorithm is provided below in Figure 2. This algorithm is the procedure performed by any node that receives the query during intra-cluster propagation, including the gateway node.

The process starts when a node receives the query. The algorithm performs a sequence of "jumps," from node to node, within the scope of a cluster. "Jump" refers to the action of sending a query from one node to another. The algorithm is parameterized with a maximum number of jumps; the role of this parameter is to avoid the unstopped propagation of the query. Since the algorithm is executed in parallel by several recipient nodes, there cannot be a global maximal number of jumps that can be tracked. Instead, the maximal number of jumps is computed along a single path, i.e., every time the query is forwarded to a node, the current number of jumps is incremented by 1. When a node sends a query to a neighbor node, it also sends the current value of the number of jumps along that path. If a node receives a query but the max number of jumps along this path is reached, it stops the local propagation. Meanwhile, the propagation may continue along other paths.

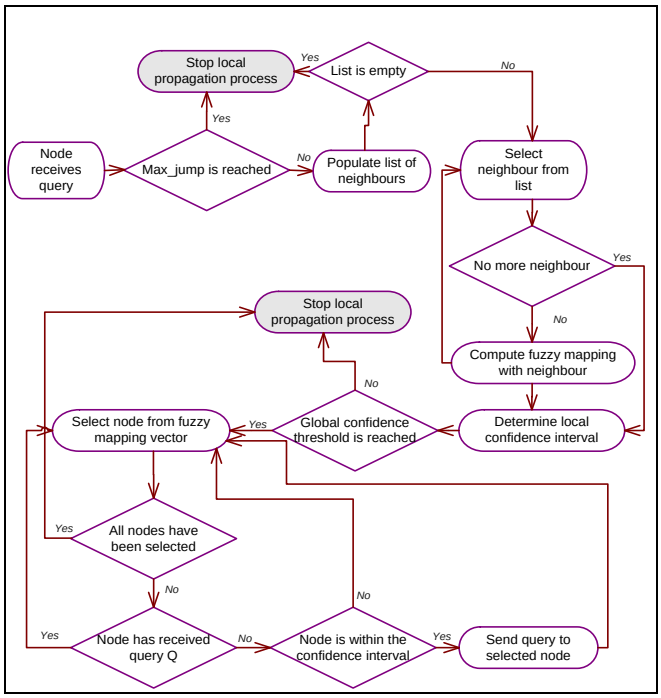


Figure 2. Fuzzy intra-cluster propagation algorithm

If the maximum number of jumps is not reached, the current recipient node creates a list of neighbor nodes. For each neighbor node, the recipient node computes a fuzzy mapping between the query and the neighbor node's metadata, which is composed of two components: a semantic relation *r* and a fuzzy inclusion value *f* (details on how the semantic relation *r* and the fuzzy inclusion value *f* are computed are given in Section V). As a result, the recipient node obtains a fuzzy mapping vector:

$$V = ((f_1, r_1), (f_2, r_2), \ldots (f_n, r_n)).$$

To determine which neighbor node(s) will be selected as new query recipient, three conditions must be verified:

- the fuzzy semantic relation *r* must be one of the type(s) selected by the user (among the possible semantic relations listed in Table 1 and presented in Section V);
- the fuzzy inclusion must fall within a local confidence interval, and
- the fuzzy inclusion must meet a global user-defined threshold.

The fact that the user can select both quantitative criteria (the fuzzy inclusion thresholds) and a qualitative criterion (the semantic relation *r*) to restrict the nodes that can be selected as query recipients gives more flexibility to the approach and makes it more adaptable to the user's needs. For example, if the user specifies that the semantic relation between the query and the neighbor node's metadata must be "contains," it means that the user accepts to receive data more specific than the needs expressed in the query. Conversely, if the user select "contained in" as a semantic relation, its means he or she accept to receive data more general than the needs expressed in the query.

The local confidence interval is a percentage of the highest values in *V*. More specifically, the local confidence interval is the interval of fuzzy inclusion values that contains *x* percent of the highest values of fuzzy inclusion in V, where *x* is a threshold that can be user-defined. For example, if *x* = 20, it means that the interval will contain 20 percent of the elements in *V*, with these 20 percent elements being the highest possible. Therefore, the smaller *x* is, the more selective is the algorithm. In the experiment presented in Section VI, we have selected *x* = 20, since the ability of the algorithm to forward the query to relevant nodes was optimal using this threshold for the given data set.

This interval is local because for every node, a different interval is determined dynamically at run-time. The purpose of having both local and global threshold is to deal with variation of fuzzy inclusion values within the network. To ensure that no node answers twice the same query, the query is given a unique identifier stored by nodes who received it. If a node receives a query it had already forwarded, it will stop the local propagation process (the propagation may continue along other paths).

V. FUZZY SEMANTIC MAPPING MECHANISM

In this section, we present the fuzzy semantic mapping mechanism that supports the query propagation process presented in the previous section. The fuzzy semantic mapping mechanism, which produces both qualitative and quantitative relations, was introduced in Bakillah and Mostafavi [1]; however, in this paper we extend it to include the cases of discrete but also continuous properties. Some papers on fuzzy ontology mapping have already been published, for example, [37][38]. However, these approaches have limited expressivity. For example, [37] focus on finding subsumption relations between concepts of fuzzy ontologies, while our fuzzy semantic mapping framework provides 9 possible mapping relations. [38] do not address the comparison of fuzzy continuous ranges of values for properties, while in this paper we integrate measures for both discrete and continuous properties.

In this paper, we assume that the metadata on sensors is formalized in an ontological format. An ontology is usually defined as a set of concepts (or classes) that represent entities of the domain of discourse, relations and/or properties, and axioms, which are statements that are true within that domain of discourse [39]. We follow a similar approach to define the fuzzy geospatial ontology. However, in the fuzzy ontology, we consider that the membership degree of a property or relation in the definition of a concept can be quantified. In a crisp ontology, the membership degree of a property or relation into the definition of a concept is always one or zero. This means that either a concept has that property; or it does

not have it. In the fuzzy ontology, this membership degree varies between zero and one, to indicate partial membership. Therefore, in a fuzzy ontology, concepts do not have a crisp definition.

We define the fuzzy geospatial ontology as a 5-tuple: $O = \{C, R, P, D, rel, prop\}$, where:

- C is a set of concepts, which are abstractions of entities of the domain of discourse;
- R is a set of relations;
- P is a set of properties for concepts;
- D is a set of possible values for properties in P, called range of properties;
- rel: $[R \rightarrow C \times C] \rightarrow [0, 1]$ is a fuzzy function that specifies the fuzzy relation that holds between two concepts;
- $prop$: $[P \rightarrow C \times D] \rightarrow [0, 1]$ is a fuzzy function that specifies the fuzzy relation between a concept and a subset of D. D is therefore a fuzzy range of values.

The set of relations R includes spatial relations such as "Is_located_at," which indicates the location of an instance of the concept, and other topological, directional and orientation spatial relations, which can be fuzzy. Therefore, in this paper we assume that either the query can contain a fuzzy property (e.g., "find sensors monitoring temperature close to point A," with point A being defined with a fuzzy function such as in Figure 4), or the sensor itself can be defined by a fuzzy property (its position in space being fuzzy), or both. The fuzzy semantic mapping mechanism takes into account these three cases, since fuzzy sets theory also include the case of crisp sets (where membership degree can only be 1 or 0).

For the purpose of our approach, we define a concept with a conjunction of a set of axioms A_C, where each axiom is a fuzzy relation or property that defines the concept:

$$C = A1 \sqcap A2 \sqcap \ldots. \sqcap An.$$

We use the term axiom, which is usually employed to refer to the whole expression that defines a concept, because a concept could also be defined by one feature (property or relation).

The idea of the fuzzy semantic mapping mechanism is to use fuzzy logics to first determine the fuzzy inclusion of a concept into another concept from a different ontology (or, in the case of fuzzy propagation, the fuzzy inclusion of the query concept into another concept describing the semantics of sensor data), based on the fuzzy inclusion of each axiom of the first concept into axioms of the second concept. Then, fuzzy predicates, which value depends on the fuzzy inclusion, are used to infer the semantic relation between the two concepts.

Let two concepts C and C' be defined as follows:

$$C = A_1 \sqcap A_2 \sqcap \ldots. \sqcap A_n$$

$$C' = A_1' \sqcap A_2' \sqcap \ldots. \sqcap A_m'.$$

We define the fuzzy semantic mapping between C and C' as follows:

Definition (fuzzy semantic mapping) A fuzzy semantic mapping m^C between C and C' is a tuple $m^C = <C, C', rel(C, C'), \mu(C, C')>$, where rel is a semantic relation between C and C', and $\mu(C, C')$ is the fuzzy inclusion of C into C'.

We define the fuzzy inclusion as the membership degree of a concept in another. This means that when the value of the fuzzy inclusion is 1, the first concept is entirely included in the second concept; when it is zero, no axiom of the first concept intersects with axioms of the second. The fuzzy inclusion of C into C' is denoted with $\mu(C, C')$:

$$\mu(C,C') = \frac{\sum_{A \in (A_1,\ldots,A_n,A_1',\ldots A_m')} \min(\mu_C(A), \mu_{C'}(A))}{\sum_{A \in (A_1,\ldots,A_n,A_1',\ldots A_m')} \mu_C(A)}, \quad (1)$$

where $\mu_C(A)$ is the membership degree of axiom A in concept C. We know that this membership degree comes from the definition of the concept in the fuzzy geospatial ontology. Now there are two cases to consider: either the axioms are formed with properties with discrete range of values, or axioms are formed with properties with continuous range of values (i.e., a fuzzy function such as in Figure 4). In each case, the fuzzy inclusion must be computed using different formulas.

First, we explain how the fuzzy inclusion of C into C' defined in (1) is computed in the case of properties with discrete or continuous range of values. Secondly, we explain how the semantic relation rel between C and C' is determined using the fuzzy inclusion value.

A. Fuzzy Inclusion: The Discrete Case

Let A: $<r.D>$ and A': $<r'.D'>$ be two axioms, where D and D' are discrete fuzzy ranges of values (e.g., temperature = low, average, or high). For example, <temperature.((0.2, low);(0.8, average))> represents the partial membership of temperature value into the set of low and average temperature intervals.

To compute (1), which relies on the membership of axiom A in concept C', and where axiom A of concept C might not be already in the definition of the concept C', we need the membership of axiom A in axiom A' of C'. The membership degree of A into A' is determined by the Zadeh conjunction for fuzzy sets:

$$\mu(A, A') = \min(\mu(D, D'), \mu(r, r')) . \quad (2)$$

Generally, the function $\mu(X1, X2)$ over any fuzzy sets $X1$, $X2$ is defined as follows, using the fuzzy implication principle of fuzzy logics [33]:

$$\mu(X1, X2) = \inf_{x \in X1 \cup X2} (\mu_{X1}(x) \Rightarrow_f \mu_{X2}(x)), \quad (3)$$

where $\Rightarrow_f$ is a fuzzy implication operator from [0,1] into [0,1], and x is any element belonging to *X1* and/or *X2*. There are several definitions for the fuzzy implication operator (including Gödel, Gogen and Lukasiewicz fuzzy implications, see [33]). We use Lukasiewicz fuzzy implication because of its superior flexibility, which is defined as follow:

$$\mu_{X1}(x) \Rightarrow_L \mu_{X2}(x) = \begin{cases} 1 & \text{if } \mu_{X1}(x) \le \mu_{X2}(x). \\ 1 - \mu_{X1}(x) + \mu_{X2}(x) & \text{otherwise} \end{cases} \quad (4)$$

Now, we need to adapt formulas (3) and (4) to compute μ(*D*, *D*') and μ(*r*, *r*'). For example, consider the problem of computing μ(*D*, *D*'). Consider also that c_i' is an element of the fuzzy set *D*'. We see from (4) that we need to know the membership degree of elements of *D*' into *D* ($\mu_D(c_i')$), and vice-versa. However, this membership degree is not readily available, because elements of *D*' are not necessarily included in *D*. In other words, all we have is the membership degree of an element into the set to which it initially belongs. Nevertheless, it is possible to compute $\mu_D(c_i')$ if we know the membership degree of c_i' into elements of *D* (denoted with c_j). To do so, we use the Lukasiewicz fuzzy composition operator, denoted with the symbol $\otimes_L$, and which determines the membership of a first element c_i' in a set *D*, knowing the membership degree of c_i' in c_j and the membership degree of c_j in *D* (Figure 3). The symbol *c* is used to indicate an element of the range of values of a property or a relation of the fuzzy geospatial ontology.

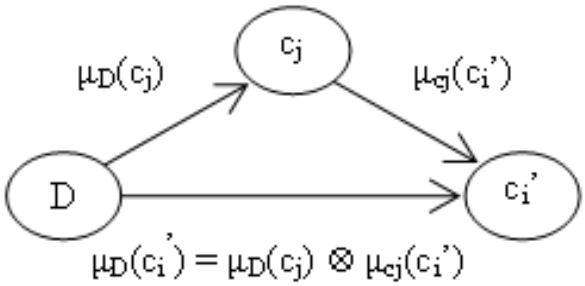


Figure 3. Illustration of the Lukasiewicz fuzzy composition principle

According to this principle, the membership degree of c_i' in *D* writes as:

$$\mu_D(c_i') = \sum_j \mu_D(c_j) \otimes_L \mu_{c_j}(c_i'), \ \forall j \left| (\neg c_j \perp c_i), \right. \quad (5)$$

where

$$\mu_D(c_j) \otimes_L \mu_{c_j}(c_i') = \max(\mu_D(c_j) + \mu_{c_j}(c_i') - 1, 0), \quad (6)$$

according to Lukasiewicz's definition of the fuzzy composition operator.

To determine $\mu_{cj}(c_i')$, which is the membership degree of an element c_i' of a range of values in an element c_j of another range of values, we have developed a fuzzy membership degree measure. This measure is based on the relative position of c_j and c_i' in an upper-level ontology *O*. An appropriate ontology for this task is a domain-independent, largely recognized lexical base, such as WordNet. However, other specialized upper-level ontologies might be more useful, depending on the domain of application. Of note however is that the chosen upper-level ontology should be structured with is-a relations. This is because is-a relations allow to identify inclusion relations between elements of the ontology, which allows to derive membership degrees. We note that using such external resource allows to deal with the terminological heterogeneity that characterizes the metadata of sensors produced by different organizations. Let $<_O$ be a hierarchical, is-a relation between terms *t* in *O*, such that $t <_O t'$ means that *t* is more specific (less general) than *t*'. Let $P(c_j, c_i')$ be the path relating c_j to c_i' in *O*, according to this hierarchy: $P(c_j, c_i') = \{c_j, t1, t2, \ldots c_i'\}$ so that *t*1, *t*2, … is the ordered set of nodes (representing terms) from c_j to c_i' in *O*. Let $d(t_k)$ be the set of descendants of a node t_k in *O*. We define $\mu_{cj}(c_i')$ as follows:

$$\mu_{c_j}(c_i') = \begin{cases} 1 & \text{if } c_i' \prec c_j \\ \dfrac{1}{\prod_{\forall t_k \in P(c_j, c_i')} |d(t_k)|} & \text{if } c_i' \succ c_j \\ 0 & \text{otherwise} \end{cases}. \quad (7)$$

This equation means that when c_i' is more specific than c_j, it is entirely included in c_j, and when c_i' is more general than c_j, $\mu_{cj}(c_i')$ decreases with the number of descendants of its subsumers. Replacing results of (7) in (6), we obtain the membership of each element of the fuzzy range *D*' in *D*, which, in turn, allows to determine μ(*D*, *D*') with (3). Equation (7) is also used to determine μ(*r*, *r*'), so these results can be replaced in (3).

The fuzzy inclusion values between the axiom also allows to determine the semantic relation between these axioms. From the fuzzy inclusion given in (2), we obtain the semantic relation between the axioms, *rel*(*A*, *A*'), using the following rules, which are derived from the fuzzy set relationship definitions:

(R1) $A \equiv A' \Leftrightarrow \mu(A, A') = 1 \ \wedge \ \mu(A', A) = 1$

(R2) $A \sqsubseteq A' \Leftrightarrow \mu(A, A') = 1 \ \wedge \ \mu(A', A) < 1$

(R3) $A \sqsupseteq A' \Leftrightarrow \mu(A, A') < 1 \ \wedge \ \mu(A', A) = 1$

(R4) $A \sqcap A' \Leftrightarrow 0 < \mu(A, A') < 1 \ \wedge \ 0 < \mu(A', A) < 1$

(R5) $A \perp A' \Leftrightarrow \mu(A, A') = 0 \wedge \mu(A', A) = 0$.

Semantic relations between the axioms will enable to determine the semantic relation between the concepts that they compose. Before we show how this can be done (in Section C), we present the case of fuzzy inclusion between properties with continuous ranges of values.

B. Fuzzy Inclusion: The Continuous Case

Properties can have continuous fuzzy ranges of value described by fuzzy membership functions. Their general form is *A*: <*p*.*f*> and *A*': <*p*'.*f*'>, where *p* and *p*' are properties and *f* and *f*' are fuzzy continuous functions. For example, Figure 4 shows the comparison of two fuzzy membership functions describing fuzzy spatial regions *sr*' and *sr*' (e.g., fuzzy spatial location targeted by the query and fuzzy spatial area of measurement of the sensor). Such function represents the uncertainty bounds for a class of fuzzy spatial regions.

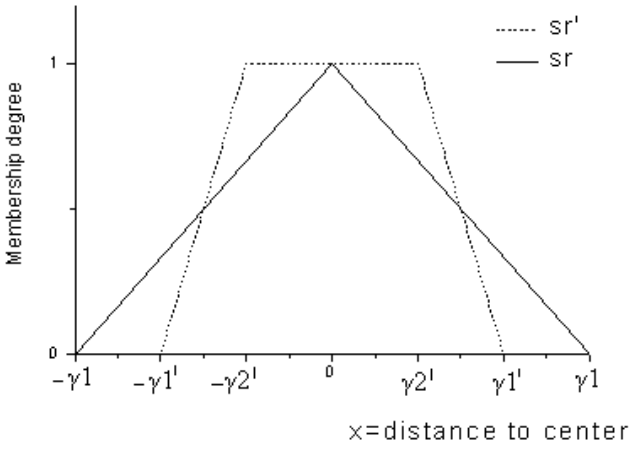


Figure 4. Example of fuzzy functions for defining a fuzzy point (geographical location)

The choice of the fuzzy function to represent the range of a property depends on the characteristics specific to the sensor, especially how the accuracy of the measurement changes in space. For example, in Figure 4, *sr* is a triangular fuzzy function; the membership degree of a point in space into the sensor's area of measurement is maximal only at a single, punctual location (*x*=0). As the distance to this single location increases, the membership degree decreases linearly and symmetrically. Such fuzzy function can be suitable to represent the area of measurement of a sensor that monitors the temperature at a certain fuzzy point, for example. Meanwhile, *sr*' represents a fuzzy trapezoidal function where the membership degree of a point in space into the sensor's area of measurement is maximal inside a given radius. Outside this radius, the membership degree decreases linearly (and more sharply than in the given triangular function). Such fuzzy function may be more suitable to describe a sensor that can detect movement within a given circular area, for example. We can also see that the slope depends on how precise the measurement is in space and therefore depends on the sensor's characteristics. Other common fuzzy functions are presented in [40].

The membership degree *A* into *A*' is computed with (2), except that the membership of *f* into *f*' is not computed with (3), which is applicable only to discrete fuzzy sets. Instead, we need to study inclusion measures for continuous fuzzy sets. A review of similarity and inclusion measures for fuzzy sets is presented in [41]. Notably, the measure for erosion of fuzzy sets by [42] is presented as a suitable measure to measure fuzzy inclusion for finite sets. Since no measurement domain of sensors can be infinite, a fuzzy inclusion for finite sets is appropriate. According to this measure, the membership of *f* into *f*' can be computed with the following function:

$$\mu(f, f') = \int_0^1 \inf_{x \in f^\alpha} \mu_{f'}(x) d\alpha \,, \qquad (8)$$

where *x* is an element of the universe of discourse (or of the union of the domains of *f* and *f*'), and f^α is called the α-cut of f, which is the binary set with defined as follows:

$$f^\alpha(x) = \begin{cases} 0 \text{ if } \mu_f(x) \prec \alpha \\ 1 \text{ if } \mu_f(x) \geq \alpha \end{cases} \qquad (9)$$

Note that this approach is used not only for spatial or temporal properties, but also for the case of thematic property axioms with fuzzy continuous ranges of value, for example A: <HasWindSpeed.Low>, where low is a continuous fuzzy range of values over the values of wind speed.

C. Semantic Relations

In order to determine the semantic relation between the query concept and a concept describing semantics of sensor data, we have defined a set of three predicates. Predicates are measurements which values are qualitative; they are used to determine whether a semantic relation between two concepts is true. The semantic relations between two concepts are qualitative relations among the following: equivalence, contains, contained in, partial symmetric-containment, partial left-containment, partial-right containment, strong overlap, weak overlap, and disjoint (as listed in Table 1). The semantic relation is determined by the following expression:

$$rel(C, C') = I(A_C, A_{C'}) \otimes_{\Pr} C(A_C, A_{C'}) \otimes_{\Pr} CI(A_C, A_{C'}), \qquad (10)$$

where $I(A_C, A_{C'})$, $C(A_C, A_{C'})$ and $CI(A_C, A_{C'})$ are three predicates that respectively evaluate the following:

- $I(A_C, A_{C'})$ predicate evaluates the intersection of axioms of the concept C with axioms of C';
- $C(A_C, A_{C'})$ predicate evaluates the inclusion of axioms of C' in axioms of C, and
- $CI(A_C, A_{C'})$ predicate evaluates the inclusion of axioms of C in axioms of C'.

The $\otimes_{Pr}$ symbol in (10) is a composition operator. Its function is to give the semantic relation between C and C', based on the value of the three predicates.

For any predicate Pr, the possible values of Pr are:

- B value, if for all axioms of C there is an axiom of C' that verifies predicate Pr, and vice-versa. For example, $I(A_C, A_{C'}) = B$ if for all axioms in A_C, there is an axiom in $A_{C'}$ that intersects this axiom (as determined by rules R1 to R5 defined in the previous section), and vice-versa;
- S value, if there exist some axioms of C and axioms of C' that verify predicate Pr, but not all;
- N value, if there exists no axiom of C and C' that verifies predicate Pr.

These principles for determining the value of a predicate are formally expressed as follows (where logic symbols are $\forall$ (for all), $\exists$ (there exists) $\perp$ (disjoint) and $\neg$ (negation)):

$$I(C,C') = \begin{cases} B & \forall i \exists j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0) \wedge \\ & \exists i \forall j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0) \\ S & \exists i \exists j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0) \wedge \\ & \neg[\forall i \exists j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0) \wedge \\ & \exists i \forall j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0)] \\ N & \neg \exists i \exists j, rel(A_i, A'_j) \neq \perp \wedge \mu(A_i, A'_j) \neq 0) \end{cases}$$

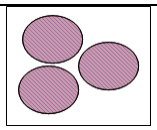
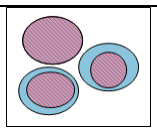
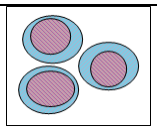
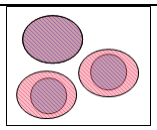
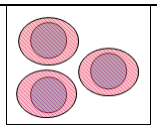
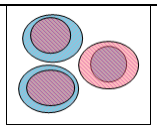
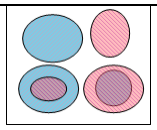
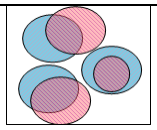
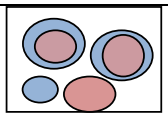
$$C(C,C') = \begin{cases} B & \forall i \exists j, rel(A_i, A_j') \in \{\equiv, \supseteq\} \wedge \mu(A_i, A_j') \neq 0 \\ S & \exists i \exists j, rel(A_i, A_j') \in \{\equiv, \supseteq\} \wedge \mu(A_i, A_j') \neq 0 \wedge \\ & \neg \forall i \exists j, rel(A_i, A_j') \in \{\equiv, \supseteq\} \wedge \mu(A_i, A_j') \neq 0 \\ N & \neg \exists i \exists j, rel(A_i, A_j') \in \{\equiv, \supseteq\} \wedge \mu(A_i, A_j') \neq 0 \end{cases}$$

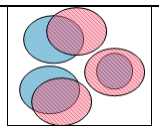
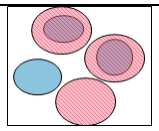
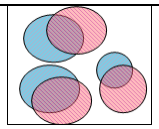
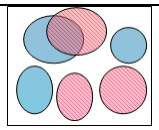
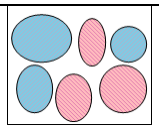
$$CI(C,C') = \begin{cases} B & \forall i \exists j, rel(A_i, A_j') \in \{\equiv, \subseteq\} \wedge \mu(A_i, A_j') \neq 0 \\ S & \exists i \exists j, rel(A_i, A_j') \in \{\equiv, \subseteq\} \wedge \mu(A_i, A_j') \neq 0 \wedge \\ & \neg \forall i \exists j, rel(A_i, A_j') \in \{\equiv, \subseteq\} \wedge \mu(A_i, A_j') \neq 0 \\ N & \neg \exists i \exists j, rel(A_i, A_j') \in \{\equiv, \subseteq\} \wedge \mu(A_i, A_j') \neq 0 \end{cases}$$

For C and C', the domain of quantifiers i and j is respectively $i \in \{1,\ldots, n\}$ and $j \in \{1,\ldots, m\}$.

As for the composition operator $\otimes_{Pr}$, it takes as input the value for the three predicates for C and C', and returns the semantic relation between C and C', according to the 14 possible combinations of predicate values identified in Table 1. This table defines the $\otimes_{Pr}$ operator: each combination of values for the three predicates is associated with a resulting semantic relation. For example, C (semantically) contains C' if $I(A_C, A_{C'}) = B$, $C(A_C, A_{C'}) = B$ and $CI(A_C, A_{C'}) = S$ (second line of Table 1). In the associated illustrations, blue sets represent axioms of C, and red sets axioms of C'.

TABLE I. SEMANTIC RELATIONS IN FUNCTION OF THE COMBINATION OF PREDICATE VALUES ($\otimes_{Pr}$ OPERATOR)

Semantic relationship (C, C')	Value of $I(A_C, A_{C'})$	Value of $C(A_C, A_{C'})$	Value of $CI(A_C, A_{C'})$	Representation
1. Equivalence	B	B	B	
2. Contains	B	B	S	
	B	B	N	
3. Contained In	B	S	B	
	B	N	B	
4. Partial S-Containment (S=Symetric)	B	S	S	
	S	S	S	
5. Partial L-Containment (L-LEFT)	B	S	N	
	S	S	N	

6. Partial R-containment (R=RIGHT)	B	N	S	
	S	N	S	
7. Strong Overlap	B	N	N	
8. Weak Overlap	S	N	N	
9. Disjoint	N	N	N	

This fuzzy semantic mapping mechanism describes how the fuzzy inclusion and semantic relation can be computed between a query concept and semantics of sensor data, and therefore supports fuzzy query propagation. It is worth noting that the approach requires the user to formulate within its query a fuzzy function for the fuzzy properties, which might not be straightforward for users who are not familiar with fuzzy set theory. Therefore, we note that further work is required to develop a friendly interface to help capture the fuzziness in user's queries in an easier fashion. Similarly, the approach requires that the fuzziness of sensor data be formally described and available within sensor metadata. In this respect, we note that several proposals have already been made for the development of Fuzzy Description Logics (DL) [40], DL being the underlying formalism of OWL, the W3C-recommended language for the Semantic Web [43].

VI. EVALUATION

In this section, we present the evaluation of the fuzzy query propagation approach. The presented evaluation is based on the comparison of the approach with the flooding algorithm, which consists in flooding the network through all available communication channels between sensors. We also compare the crisp version of the algorithm with the fuzzy version to verify whether the fuzzy algorithm helps to find more relevant sensors than the crisp version. Finally, to further investigate the behavior of the algorithm, we compare the results using different fuzzy inclusion thresholds as criteria to select query recipient nodes.

The approach was implemented as a simulation in Java (Eclipse 3.4, JDK 1.6) with a maximum of 20,000 nodes. Nodes were randomly assigned metadata using a set of metadata into which variations were randomly introduced. The original metadata was obtained from the SensorML descriptions available on the Geospatial Cyberinfrastructure for Environmental Sensing platform (GeoCENS), an online platform that enables simplified searching, storing and sharing of environmental and other georeferenced data [2], to which we have added fuzzy membership functions on their location and some thematic attributes (e.g., temperature, precipitations and soil moisture) for the purpose of the simulation.

The simulations performed were compared in terms of the rate of dissemination of the query to the relevant sensor nodes. The approach is efficient if the least sensor nodes are sent messages for a maximum of relevant nodes being reached and identified as query recipients. The rate of dissemination compares the percentage of relevant nodes that were selected as query recipients (vertical axis) versus the number of sensor nodes that were reached (i.e., that received the query message) (horizontal axis). Therefore, we are not only evaluating the ability of the algorithm to propagate the query while reducing energy consumption, but also the ability to find the best path to maximize the recall and accuracy. The relevant nodes with respect to a query were identified manually during the setting of the simulation and used as authoritative result for the evaluation of the approach.

Figure 5 shows the assessment of the rate of dissemination for the flooding, crisp, and fuzzy algorithms, tested with a fuzzy inclusion threshold of 0,40.

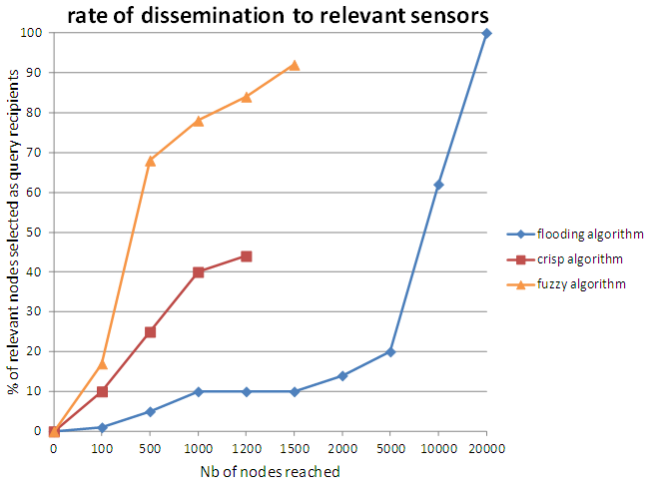


Figure 5. Rate of dissemination using flooding, crisp or fuzzy propagation algorithm

The flooding algorithm, because it reaches all nodes of the network, is able to achieve a 100 percent recall of relevant sensor nodes. But this is only at the very high cost of sending messages to all nodes of the network, which is not appropriate in an environment where the energy of sensors must be saved since it is not guaranteed that sensors

are easily accessible and can be replaced or their energy source renewed; for example some sensors are buried to measure soil moisture, while others are underwater to measure water temperature, etc. Meanwhile, the crisp algorithm can only achieve a 43 percent recall of relevant nodes. This is because the crisp query is very restrictive in comparison with the fuzzy query. While the fuzzy algorithm is more costly that the crisp algorithm (20 percent more nodes received a query message), its performance counterbalances this cost since the recall of relevant nodes reaches over 90 percent.

Figure 6 shows the rates of query dissemination to relevant sensors for different values of the fuzzy inclusion threshold (0,20, 0,40, 0,60 and 0,80).

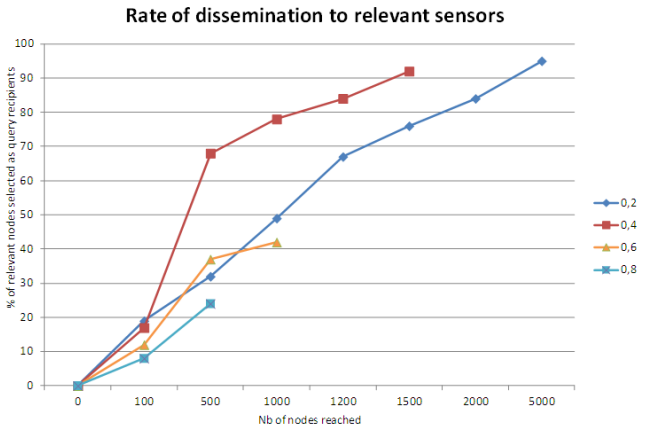


Figure 6. Rate of dissemination using various fuzzy inclusion thresholds

When the algorithm is set with a 0,20 or 0,40 fuzzy inclusion threshold, the difference in the recall of relevant nodes is very slight, suggesting that even low fuzzy inclusion between 0,20 and 0,40 might be sufficient to indicate relevance.

However, with a low threshold, a significant number of sensor nodes that are not relevant are accessed in comparison with the 0,40 threshold. With the 0,60 and 0,80 thresholds, an important percentage of relevant nodes are missed and the query propagation is stopped after reaching a smaller number of nodes. Although this study does not demonstrate which threshold is appropriate at all times, since this is likely to depend on the data being used, this demonstrates that the choice of the fuzzy inclusion threshold is a determining factor influencing the efficiency of the approach. Therefore, a testing phase with sample network is necessary to establish the more relevant threshold.

VII. Conclusion and Future Work

In the geospatial domain, it is essential to consider the uncertainty and fuzziness of geospatial phenomena. In a previous paper, we had presented an approach for fuzzy semantic mapping of fuzzy geospatial ontologies [6]. In this paper, we have demonstrated one of the possible applications of this approach through incorporating it into a new approach for fuzzy query propagation in sensor networks.

Sensors are devices intended to monitor environmental conditions, and they can be interconnected through so-called sensor networks. Due to energy, processing and memory limitations pertaining to their size, sensors of a network cannot be all reached by an application. They must rather be queried and their data retrieved through intermediary sensor nodes of the network. This situation creates the need for query propagation mechanisms that are efficient in terms of cost, but that are also able to retrieve requested data. At the same time, we believe that the fuzziness of query and sensor data must be taken into account in query propagation to improve the ability to retrieve relevant data. This created the motivation for the fuzzy query propagation approach that has been proposed in this paper. The fuzzy query propagation approach comprises a first propagation step towards relevant clusters of sensors, therefore varying the sensor nodes that will have to redistribute the query; it is followed by an intra-cluster query propagation phase. In both phases, the fuzzy semantic mapping mechanism is used to select query recipients. The experiments that were conducted show that in comparison with a crisp approach, taking into account the fuzziness indeed improves the recall of relevant data while avoiding the increase of propagation cost. We have also noted that one challenge or limitation raised by our research is related to the impact on the performance of the approach of some parameters of the proposed algorithm, including the fuzziness threshold being chosen to select query recipients. Therefore, further research is required to investigate avenues for helping the user to select the appropriate threshold in a user-friendly fashion.

Among future work being uncovered by this study, we plan to investigate the role of such fuzzy query propagation approach into the so-called semantic enablement of Spatial Data Infrastructures (SDIs). Because the objective of SDIs is to support the exchange of heterogeneous data and information among various providers and users, future research on SDIs will aim at integrating access to sensor networks through SDIs. Therefore, we foresee that future work on how to integrate fuzzy query propagation as a service into SDIs will be useful. Semantic-based query propagation strategies such as provided in this paper can be adapted to SDIs and coordinated with catalogue services so that the user can, through a single interface, search for either data from web services registered in centralized catalogues or data from dynamic networks made accessible through SDIs.

Acknowledgment

This research was made possible by an operating grant from Microsoft Research and Alberta Innovate Technology Future Natural Sciences.

REFERENCES

[1] M. Bakillah and M.A. Mostafavi, "A Fuzzy Logic Semantic Mapping Approach for Fuzzy Geospatial Ontologies," Proc. of SEMAPRO 2011, Lisbon, Portugal, November 2011, pp. 21-28.

[2] http://www.geocens.ca/ Access date: 03.06.2012

[3] L. J. G. Villalba, A. L. S. Orozco, A. T. Cabrera, and C. J. B. Abbas, "Routing Protocols in Wireless Sensor Networks," Sensors, vol. 9, issue 11, 2009, pp. 8399-8421. doi: 10.3390/s91108399. Access date: 03.06.2012

[4] V.B. Robinson, "A Perspective on the Fundamentals of Fuzzy Sets and Their Use in Geographical Information Systems," Transactions in GIS, vol. 7, issue 1, 2003, pp. 3-30. doi: 10.1111/1467-9671.00127 Access date: 19.12.2012

[5] O. Ahlqvist, "Using Uncertain Conceptual Space to Translate between Land Cover Categories," International Journal of Geographical Information Science, vol. 19, issue 7, 2005, pp. 831-857. doi: 10.1080/13658810500106729 Access date: 19.12.2012

[6] C. Hudelot, J. Atif, and I. Bloch, "Fuzzy Spatial Relation Ontology for Image Interpretation," Fuzzy Sets and Systems, vol. 159, 2008, pp. 1929-1951. doi:10.1016/j.fss.2008.02.011 Access date: 19.12.2012

[7] H. Karl and A. Willig, "Protocols and Architectures for Wireless Sensor Networks," Chichester, West Sussex, UK: John Wiley & Sons, 2005. ISBN: 978-0-470-09510-2

[8] C. Liu, K. Wu, and J. Pei, "An Energy-efficient Data Collection Framework for Wireless Sensor Networks by Exploiting Spatiotemporal Correlation," IEEE Transactions on Parallel Distribution Systems, vol. 18, 2007, pp. 1010-1023. doi: 10.1109/TPDS.2007.1046 Access date: 19.12.2012

[9] B. Gedik, L. Liu, and P.S. Yu, "ASAP: An Adaptive Sampling Approach to Data Collection in Sensor Networks," IEEE Transations on Parallel Distribution Systems, vol. 18, 2007, pp. 1766-1783. doi:10.1109/TPDS.2007.1110. Access date: 19.12.2012

[10] R. Teng and B. Zhang, "On-demand Information Retrieval in Sensor Networks with Localised Query and Energy-balanced Data Collection," Sensors, vol. 11, 2011, pp. 341-361. doi: 10.3390/s110100341. Access date: 03.06.2012

[11] A. Zafeiropoulos, D.-E. Spanos, S. Arkoulis, N. Konstantinou, and N. Mitrou, "Data management in sensor networks using semantic web technologies," Data Management in Semantic Web, H. Jin, Z. Lv, Eds. Nova Science Publishers, Inc., 2009, pp. 97-118.

[12] P.B. Karp, Y. Ke, S. Nath, and S. Seshan, "IrisNet: An Architecture for a Worldwide Sensor Web," IEEE Pervasive Computing, vol. 2, 2003, pp. 22-33. doi: 10.1109/MPRV.2003.1251166. Access date: 19.12.2012

[13] L. Kulik, E. Tanin, and M. Umer, "Efficient Data Collection and Selective Queries in Sensor Networks," Proc. of 2nd International Conference on GeoSensor Networks, Boston, MA, USA, October 2006, pp. 25-44. doi: 10.1007/978-3-540-79996-2-3. Access date: 19.12.2012

[14] J. Kulik, W. Heinzelman, and H. Balakrishnan, "Negotiation-based Protocols for Disseminating Information in Wireless Sensor Networks," Wireless Networks, vol. 8, 2002, pp. 169–185. doi: 10.1023/A:1013715909417. Access date: 19.12.2012

[15] D. Braginsky and D. Estrin, "Rumor Routing Algorithm for Sensor Networks," Proc. of the First ACM International Workshop on Wireless Sensor Networks and Applications (WSNA), Atlanta, GA, USA, September, 2002, pp. 22–31. doi:10.1145/570738.570742. Access date: 19.12.2012

[16] N. Sadagopan, B. Krishnamachari, and A. Helmy, "The ACQUIRE Mechanism for Efficient Querying in Sensor Networks," Proc. of the First IEEE International Workshop on Sensor Network Protocols and Applications (SNPA), Anchorage, AK, May 2003, pp. 149–155. doi: 10.1109/SNPA.2003.1203365. Access date: 19.12.2012

[17] S. Lindsey and C.S. Raghavendra, "PEGASIS: Power-efficient Gathering in Sensor Information Systems," Proc. of the Aerospace Conference, Big Sky, MT, March, 2002, pp. 1125–1130. doi: 10.1109/AERO.2002.1035242. Access date: 19.12.2012

[18] S. Chatterjea, S. De Luigi, and P. Havinga, "DirQ: a Directed Query Dissemination Scheme for Wireless Sensor Networks," Proc. of the IASTED International Conference on Wireless Sensor Networks (WSN), Banff, Alberta, Canada, July 2006. doi:10.1109/ICPPW.2006.20. Access date: 19.12.2012

[19] K. Seada and A. Helmy, "Geographic Protocols in Sensor Networks," Technical Report 04-837, Computer Science Department, University of Southern California: San Diego, CA, USA, 2008.

[20] T. He, J.A. Stankovic, C. Lu, and T.F. Abdelzaher, "SPEED: a Stateless Protocol for Real-time Communication in Sensor Networks," Proc. of the 23rd International Conference on Distributed Computing Systems (ICDCS), Providence, RI, USA, May, 2003, pp. 46–55. doi: 10.1109/ICDCS.2003.1203451. Access date: 19.12.2012

[21] I. Stojmenovic, "Geocasting with Guaranteed Delivery in Sensor Networks," IEEE Wireless Communication Magazine, vol. 11, 2004, pp. 29-37. doi: 10.1109/MWC.2004.1368894. Access date: 19.12.2012

[22] H. Couclelis, "The Certainty of Uncertainty: GIS and the Limits of Geographic Knowledge," Transactions in GIS, vol. 7, issue 2, 2003, pp. 165-175. doi: 10.1111/1467-9671.00138. Access date: 19.12.2012

[23] J. Zhang and M. Goodchild, "Uncertainty in Geographical Information," London: Taylor & Francis, 2002.

[24] L. A. Zadeh, "Fuzzy Sets," Information and Control, vol. 8, issue 3, 1965, pp. 338–353. doi: 10.1016/S0019-9958(65)90241-X. Access date: 19.12.2012

[25] A. Hagen, "Fuzzy Set Approach to Assessing Similarity of Categorical Maps," International Journal of Geographical Information Science, vol. 17, issue 3, 2003, pp. 235–249. doi: DOI:10.1080/13658810210157822. Access date: 19.12.2012

[26] S. Swapna Kumar, M. Nanda Kumar, and V. S. Sheeba, "Fuzzy Logic based Enercy Efficient Hierarchical Clustering in Wireless Sensor Networks," International Journal of Research and Reviews in Wireless Sensor Networks, Vol. 1, No. 4, 2011, pp. 53-57. doi: 10.1109/CHUSER.2011.6163758. Access date: 19.12.2012

[27] S. A. Munir, Y. Wen Bin, R. Biao, and M. Jian, "Fuzzy Logic based Congestion Estimation for QoS in Wireless Sensor Network," Proc. of 2007 WCNC, IEEE, 2007, pp. 4339-4344. doi: 10.1109/WCNC.2007.791. Access date: 19.12.2012

[28] K. Kim and H. Suk Seo, "A Trust Model Using Fuzzy Logic in Wireless Sensor Network," Proc. of World Academy of Science, Engineering and Technology, Vol. 42, 2008, pp. 63-66.

[29] Y.-J. Wen, A. M. Agogino, and K., Goebel, "Fuzzy Validation and Fusion for Wireless Sensor Networks," Proc. of 2004 ASME International Mechanical Engineering Congress and RD&D Expo, November 13-19, 2004, Anaheim, California, USA, 2004, pp. 1-6. doi: 10.1115/IMECE2004-60964. Access date: 19.12.2012

[30] S. Hoon Chi and T. Ho Cho, "Fuzzy Logic based Propagation Limiting Method for Message Routing in Wireless Sensor Networks," Proc. of 2006 Computational Science and Its

Applications, LNCS 3983, 2006, pp. 58-67. doi: 10.1007/11751632-7. Access date: 19.12.2012

[31] T. Srinivasan, R. Chandrasekar, and V. Vijaykumar, "A Fuzzy, Energy-efficient Scheme for Data Centric Multipath Routing in Wireless Sensor Networks," Proc. of 2006 International Conference on Wireless and Optical Communications Networks, IEEE, Bangalore, India. doi: 10.1007/11751632-7. Access date: 19.12.2012

[32] M. Yusuf and T. Haider, "Energy-aware Fuzzy Routing for Wireless Sensor Networks," Proc. of the 2005 IEEE Symposium on Emerging Technologies, 17-18 Sept. 2005, IEEE, 2005, pp. 63-69. doi: 10.1109/ICET.2005.1558856. Access date: 19.12.2012

[33] P. Bosc and O. Pivert, "About Approximate Inclusion and its Axiomatization," Fuzzy Sets and Systems, vol. 157, 2006, pp. 1438-1454. doi: 10.1016/j.fss.2005.11.011. Access date: 19.12.2012

[34] W.R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "Energy-efficient Communication Protocol for Wireless Microsensor Networks," IEEE Computer Society Proc. of the 33nd Hawai International Conference on System Science, Washington, DC, USA, vol. 8, 2000, pp. 8020. doi: 10.1109/HICSS.2000.926982. Access date: 19.12.2012

[35] M. Bakillah and S. H.L. Liang, "Discovering Sensor Services with Social Network Analysis and Expanded SQWRL Querying," Proc. of W2GIS 2012, LNCS 7236, S. Di Martino, A. Peron, and T. Tezuka, Eds. Berlin Heidelberg: Springer Verlag, 2012, pp. 221-238. doi: 10.1007/978-3-642-29247-7_16. Access date: 19.12.2012

[36] M. Botts et al., "OGC Sensor Web Enablement: Overview and High Level Architecture," (OGC 07-165), Open Geospatial Consortium white paper, 2007.

[37] B. Xu, D. Kang, J. Lu, Y. Li, and J. Jiang, "Mapping Fuzzy Concepts Between Fuzzy Ontologies," Proc. of the 9th International KES Conference, Melbourne, Australia, 2005, LNCS 3683, pp. 199-205. doi: 10.1007/11553939_29. Access date: 20.12.2012

[38] S. Niwattanakul, P. Martin, M. Eboueya, and K. Khaimook, "Ontology Mapping based on Similarity Measure and Fuzzy Logic," Proc. of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, T. Bastiaens and S. Carliner (Eds.), 2007, pp. 6383-6387.

[39] P. Agarwal, "Ontological Considerations in GIScience," International Journal of Geographical Information Science, vol. 19, issue 5, 2005, pp. 501–536. doi: 10.1080/13658810500032321. Access date: 19.12.2012

[40] F. Bobillo and U. Straccia, "FuzzyDL: An Expressive Fuzzy Description Logic Reasoner," Proc. of IEEE International Conference on Fuzzy Systems, 1-6 June 2008, pp. 923-930. doi: 10.1109/FUZZY.2008.4630480. Access date: 19.12.2012

[41] M.-S. Yang and D.-C. Lin, "On Similarity and Inclusion Measures between Type-2 Fuzzy Sets with an Application to Clustering," Computers and Mathematics with Applications, vol. 57, 2009, pp. 896-907. doi:10.1016/j.camwa.2008.10.028. Access date: 20.12.2012

[42] L. Bloch, H. Maitre, "Fuzzy Mathematical Morphologies: A Comparative Study," Pattern Recognition, vol. 28, 1995 pp. 1341-1387.

[43] K. Janowicz, C. Keßler, M. Schwarz, M. Wilkes, I. Panov, M. Espeter, and B. Baeumer, "Algorithm, Implementation and Application of the SIM-DL Similarity Server," Proc. of the Second International Conference on GeoSpatial Semantics (GeoS 2007), Mexico City, Mexico, 29–30 November 2007, pp. 128–145. doi: 10.1007/978-3-540-76876-0-9. Access date: 19.12.2012

Enhancing Environment Perception for Cooperative Power Control: an Experimental Perspective

Panagiotis Spapis, George Katsikas, Konstantinos Chatzikokolakis, Roi Arapoglou, Makis Stamatelatos, and Nancy Alonistioti

Department of Informatics and Telecommunications
National and Kapodistrian University of Athens
Athens, Greece
e-mail:{pspapis, katsikas, kchatzi, k.arapoglou, makiss, nancy} @di.uoa.gr

Abstract - **Short range communications in dense residential environments enable anytime high data rate connectivity, however also pose new challenges regarding the efficient operation of network devices, related to their co-existence. These challenges mainly concern capacity requirements on the one hand and the interference effect that each device creates to its neighboring ones on the other. This paper presents a cooperative distributed algorithm for power control and interference mitigation based on ad-hoc communication of networking devices. The algorithm also incorporates learning capabilities for strengthening the situation perception of each network element. Both versions of the algorithm, the core cooperative power control, and the learning enhanced one, have been deployed in WiFi Access Points and tested in an office environment in order to showcase their applicability. The experimental results prove that the incorporation of the presented algorithms leads to significant gains both in the energy consumption and the interference mitigation at the same time.**

Keywords – co-existence; cooperative power control; interference mitigation; learning; data mining; fuzzy logic.

I. INTRODUCTION

The acute proliferation of wireless networking devices enables "anytime" and "anywhere" communications. This trend, coupled with large scale deployment of heterogeneous radio access networks in short range context, (APs, pico-cells, etc.) and in dense environments, (i.e., residential areas) imposes the need for developing mechanisms addressing issues related to co-existence in an efficient way. The capacity and energy efficiency requirements impose different constraints in the system, whereas the mentioned co-existence results in high interference levels.

In such communication environments, power control mechanisms aim at optimizing the network's capacity and coverage and at the same time at achieving interference mitigation, reducing power consumption and extending battery lifetime. The purpose is to have improved QoS for the users as well as having the optimum overall network's utility and reduced cost from the network operator's perspective. Given the two aforementioned objectives, the mechanisms should be developed following a cooperative and distributed paradigm in order to avoid selfish behaviors that lead to suboptimum solutions.

In this paper, a distributed and cooperative power control algorithm is presented and evaluated; the objective is, through power adjustment, to have an optimum tradeoff between the network elements' capacity and the interference caused to the rest of the network elements belonging in the scheme. The Cooperative Power Control (CPC) algorithm, initially described in our previous work in [1], is applicable to short-range wireless networking environments, where the network elements are able to exchange interference and power information. Moreover, the proposed solution deploys learning capabilities to the devices in order to facilitate the evaluation of the previous decisions and improve the interpretation of the environment conditions. This paper builds on the previous work and presents an extensive experiment for the validation of the proposed algorithm. More specifically, we have developed the CPC algorithm and incorporated it in WiFi APs; our solution has been used in a real life experiment, in an office network environment, which highlights the merits from its incorporation in both energy consumption and interference mitigation.

The rest of this paper is structured as follows: Section II presents proposed solutions available in the literature; Section III provides background information regarding fuzzy logic and k-Means; in Section IV, the baseline reference algorithm for cooperative power control is briefly described. Section V presents the learning-assisted algorithm, by describing the considered functionalities, the case study, and, the learning framework. Section VI describes the experimentation deployment and assumptions of the experimental analysis, whereas Section VII describes and analyses in details the experimental results. Finally, Section VIII concludes the paper.

II. RELATED WORK

The transmission power control adjustment has attracted the interest of researchers, given the benefits stemming from the introduction of power control schemes; thus several solutions have been proposed in the literature. In [2], Sun et al. propose to formulate the power control problem using a non-cooperative game; the solution converges once Nash equilibrium [3] is reached and is applicable to mobile ad-

hoc networks. The strategy for the transmission power identification is related to the Shannon capacity [4] on the one hand and the energy waste due to the caused interference on the other. In [5], the authors introduce a competitive distributed and autonomous power control algorithm for cellular communication systems. This approach is mainly focused on the downlink communication but can be easily extended to take into account both downlink and uplink. The nodes set independently their Signal to Interference Ratio (SIR) targets and rely only on local information to proceed to power adjustment. The algorithm is proven to converge to the Pareto optimal solution when the system is feasible, but diverges otherwise [6]. In [7], a cooperative game-theoretic mechanism for optimizing power control is also proposed. In this solution, issues such as network efficiency and user fairness are taken into account in order to optimize a SINR-based utility function.

The afore-described solutions are generic and focus on the transmission power control problem in general. However, specific solutions have been proposed in the literature, trying to tackle the transmission power control problem in WiFi networks. In [8], Mhatre et al. propose a power control algorithm that tries to mitigate interference in 802.11 wireless network environments, by providing a starvation-free transmission scheme based on the assignment of higher transmission power to cells that are more heavily loaded (i.e., the cells that have higher number of clients or clients with poor quality channel). This solution can be implemented in a centralized or a distributed manner; in case of the former the authors use a sampler in order to compute the optimum power vector of the AP topology by avoiding extensive signaling. However, misbehavior may occur in case power-vectors of high probability exist, as the algorithm fails to search other possible vectors. In [9], a synchronous rate and power control system implemented in IEEE 802.11 AP is introduced. Such solution provides per-link power control without adaptations or modifications to the underlying 802.11 MAC protocol, following an approach with two synchronized phases. In the former, an initial power level is identified so as to achieve admirable link performance whereas in the latter, further enhancements in the data rate and the transmission power level, based on the packet delivery rate are considered for avoiding performance degradation. The main disadvantage of this solution is the use of greedy schemes for power level allocation that cannot provide a maximum network throughput. In [10], Kowalik et al. propose the introduction of ConTCP, a power adaptation scheme that takes into consideration the links' quality. Specifically, a reference node tries to calculate the approvable power level of each incoming wireless link, based on QoS level thresholds, and informs the AP for the selected power levels; the proposed scheme tends to perform well under specific network topologies, where simultaneous transmissions occur. In [11], the authors propose a power control method which discovers the required data-rate link within the transmission range through adjusting the transmission power to corresponding levels by recursively sensing the environment; this topology information is also used for the selection of the optimal route, in case of 802.11b WiFi mesh networks. In [12], ElBatt et al. propose a power management scheme for wireless ad-hoc networks with low mobility patterns; the classical shortest path routing algorithm coupled with the identification of the optimum transmission power level is used. This approach results to small clusters of ad-hoc nodes. However, even though the cluster-based interference is reduced, retransmission of packets and increase in the whole network interference is unavoidable.

In terms of this paper, we apply a solution described in [1], aiming at power control in WiFi networks in a distributed cooperative manner. Our solution is based on and extends a cooperative power control scheme for wireless sensors [13] [14]. In the proposed approach, the CPC algorithm is applied in 802.11 WiFi networks and is also enhanced by introducing a learning scheme to strengthen the situation perception capabilities of each network element. The solution is based on a hybrid model which exploits the merits of fuzzy logic and data clustering. Compared to the rest of the afore presented solutions, the proposed and implemented one aims at maximizing network utility, which is being captured by the Shannon capacity and the interference caused to the neighboring APs. Thus, the benefit is in the overall network utility which also benefits the SINR in every node. Furthermore, given the fact that we use an adaptation mechanism for enhancing the situation perception of each network element, we ensure that the APs' configuration will be the most suitable for the context where there are placed.

III. BACKGROUND

This section provides the background for the proposed solution. The baseline algorithm is based on an objective function which uses fuzzy logic for the calculation of the weights of the parts of the equation. The learning algorithm uses the k-Means data mining technique for the adaptation of the fuzzy logic controllers. The rest of this section presents the principles of fuzzy logic and k-Means so as to create a standalone document.

A. *Fuzzy logic*

Fuzzy logic is an ideal tool for dealing with complex multi-variable problems; the nature of the decision making mechanism makes it very suitable for problems with often contradictive inputs. A fuzzy reasoner (Fig. 1) consists of three parts, namely:

- The fuzzifier, which undertakes to transform the input values (crisp values) to a degree that these inputs belong to a specific state (e.g., low, medium, high, etc.) using the input membership functions.
- The inference part, which correlates the inputs and the outputs using simple "IF...THEN..." rules. Each rule results to a specific degree of certainty for each output; these degrees then are being aggregated.

- The defuzzifier, where the outcome of the abovementioned aggregation is being mapped to the degree of a specific state that the decision maker belongs to. Several defuzzification methods exist; the most popular is the centroid one, which returns the center of gravity of the degrees of the outputs, taking into account all the rules, and is calculated using the following mathematical formula:

$$u_{COG} = \frac{\int u_i \mu_F(u_i) du}{\int \mu_F(u_i) du} \quad (1)$$

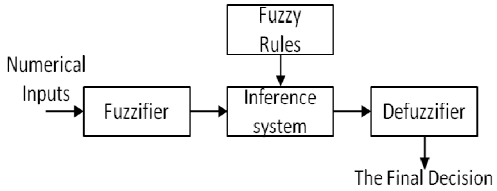


Figure 1: High level view of a fuzzy inference system

B. k-Means

k-Means is a well known data-mining clustering technique. The core idea of data clustering is to partition a set of N, d-dimensional, observations into such groups that intra-group observations exhibit minimum distances from each other (Fig. 2), while inter-group distances are maximized. k-Means [15] is based on the following objective function:

$$J = \sum_{i=1}^{c} J_i = \sum_{i=1}^{c} \left(\sum_{k, x_k \in G_i} \| x_k - c_i \| \right) \quad (2)$$

Where

- c: the number of clusters,
- G_i: the i^{th} group,
- x_k: the k^{th} vector in group J_i and represents the Euclidean distance between x_k and the cluster center c_i.

The partitioned groups are defined by using a membership matrix described by the variable U. Each element U_{ij} of this matrix equals to 1 if the specific j^{th} data point x_j belongs to cluster i, and 0 otherwise. The element U_{ij} is analyzed as follows:

$$U_{ij} = \begin{cases} 1, if \ \| x_j - c_j \|^2 \le \| x_j - c_k \|^2, \forall k \ne i \\ 0, otherwise \end{cases} \quad (3)$$

This means that x_j belongs to group i, if c_i is the closest of all centers.

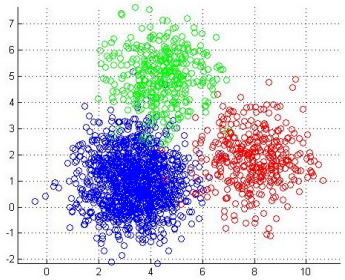


Figure 2: Visualization of k-Means clustering for three clusters

IV. Cooperative Power Control- Baseline Algorithm

The proposed CPC algorithm is based on [13] and [14]; both approaches propose a scheme for distributed interference compensation in Cognitive Radio that operates in license exempt spectrum bands, using transmission power adjustment methodologies. The solution concerns ad-hoc networks and is based on an information exchange scheme for the identification of the appropriate transmission power levels. Each independent node of the topology sets its power by considering individual information, as well as information related to the neighboring nodes. More specifically, a node sets its power level by considering its Signal to Interference plus Noise Ratio (SINR) and the interference caused to its neighbors. The main idea of this approach is to prevent users to operate in the maximum transmission power levels.

The authors assume a set of node pairs L that operate in the same frequency. The SINR for the i^{th} pair is given below [13]:

$$\gamma_i(p_i^k) = \frac{p_i^k \cdot h_{ii}}{n_o + \sum_{j \ne i} p_j^k \cdot h_{ji}} \quad (4)$$

Where

- p^k_i: transmission power for user i on channel k
- h_{ii}: link gain between i^{th} receiver and i^{th} transmitter
- n_o : noise level (equals to 10^{-2})
- p^k_j: transmission power for all other users on channel k, assuming that $j \in \{1,2,\ldots,L\}$ and $j \ne i$
- h_{ji}: link gain between i^{th} receiver and j^{th} transmitter

It is also assumed that the channel is flat-faded without shadowing effects. Since the channel is static, the only identified attenuation is the path loss h (channel attenuation or channel gain). Given that indoor urban environments are considered, the channel gain is $h_{ji} = d_{ji}^{-3}$, where d is the distance between the j^{th} transmitter and the i^{th} receiver.

The decision for the transmission power levels takes into account the negative impact (i.e., interference) of a node to its neighboring nodes. This is formalized using (5), which captures the notion of interference price; such price reflects the interference a user causes to other users within its transmission range and is given by:

$$\pi_i^k = \frac{\partial u_i(\gamma_i(p_i^k))}{\partial(\sum_{j \neq i} p_j^k \cdot h_{ji})} \quad (5)$$

Where

- $u_i(\gamma_i(p^k{}_i)=\theta_i log(\gamma_i(p^k{}_i))$: logarithmic utility function,
- θ_i: user dependent parameter.

Both of the algorithms presented in [13] and [14] are based on a tradeoff between the capacity of a user and the interference caused to the corresponding neighborhood. This balance is being captured by the following objective function:

$$u_i(\gamma_i(p_i^k)) - \alpha \cdot p_i^k \sum_{j \neq i} \pi_j^k \cdot h_{ji} \quad (6)$$

The first part indicates a relation to the Shannon capacity for the corresponding user, while the second part captures the negative impact in terms of interference prices that a user causes to its neighborhood. The *a* factor is introduced so as to capture uncertainties in the network; these uncertainties reflect the precision of the received and compiled information of each network element regarding the interference price which should have been available by the node's neighbors. This is related to the fact that once a network element adjusts its transmission power, it informs its neighbors in an ad-hoc manner. This implies that even though a network element has collected information from all of its neighbors in order to adjust its transmission, the gathered data could be obsolete and, as a consequence, they will not capture neighborhood's current state. The obsolescence of the interference prices is related to the update interval (i.e., the periodic update) of each network element. In [13], α is set in a static manner as 25%. In [14], a fuzzy reasoner is introduced in order to identify, in a more dynamic way, uncertainties in the network based on the network's status; the inputs (number of users, mobility, update interval) of the fuzzy reasoner capture the volatile nature of the ad-hoc network, whereas the output of the fuzzy reasoner is the *Interference Weight*. The *a* factor is defined as 1/β *Interference Weight* + 1 (β has the maximum value of the *Interference Weight*).

The algorithm consists of three steps, namely, the initialization, the power update and the interference price update. The former is related to the assignment of initial valid transmission power and interference price values. The second part concerns the transmission power update based on the interference prices each node receives from its neighbors. Finally, the interference price update captures the communication of its interference prices to the neighborhood, by every network node. The second and the third steps are asynchronously repeated until the algorithm reaches a steady state (i.e., a state where every network element has the same transmission power for two consecutive time iterations).

The main deficiency of the afore-described scheme is related to the static perception of the environment (i.e., *a* factor that captures the network's dynamics). Even in the case where the fuzzy reasoner is used for capturing the uncertainties in the network, the environment interpretation model (i.e., membership functions of the fuzzy reasoner) is static. More specifically, in the latter case, the environment interpretation is based on expert's knowledge and is induced to the network elements by its input membership functions. This implies that all network elements with the same configuration have the same situation perception as well. Moreover, it would be a major benefit for the network administrators to enable network elements to evolve the way they interpret their environment; this could be achieved by changing the shape of the input membership functions. In order to tackle the static definition of the situation perception, we propose a feedback-based learning scheme that evaluates how the network performed after a transmission power adjustment, in terms of the interference prices.

V. LEARNING ENHANCED COOPERATIVE POWER CONTROL FRAMEWORK

In our previous work in [1], we have proposed the application of the algorithms introduced in [13] and [14] in a completely new application area, that of WiFi Access points; the cooperative power control among the network elements is the objective of this algorithm in order to maximize the network's utility. More specifically, we suggest that the WiFi APs should cooperate in order to minimize the caused interference, by adjusting their transmission power and at the same time having the optimum transmission power based on the Shannon capacity.

In terms of this paper the learning enhanced CPC is being presented and evaluated in a real life experiment. In this section the functionalities that should be incorporated in the CPC enabled network elements (i.e., in this case WiFi APs) are described. Then, the case under investigation, where the modified CPC is applied is being presented along with the learning algorithm used for the adaptation of the situation perception of the CPC.

A. Functional Architecture

In order to deploy the CPC in the considered environment, network elements should be enhanced with a set of software modules namely "Power Control", the "Learning", the "Memory", the "CPC communication", the "Control Engine" and the "Monitoring". Fig. 3 presents the functional architecture of the software implementation of the CPC.

Each software module provides a set of functionalities in order to enable the instantiation of the CPC in WiFi APs; more specifically:

- The "Power Control" incorporates the functionalities for the calculation of the metrics (interference prices) and the objective function that each network element has to maximize. Furthermore, this part of the mechanism implements the fuzzy logic reasoner

for the calculation of the *Interference Weight* and the α factor,

- The "Learning" part incorporates the learning mechanism for enhancing the network element's situation perception.
- The "Memory" contains all the information required for the CPC; this information may be local and related to the AP under consideration (ex. TxPower, SINR, local IPs and MACs, etc.), or related to neighboring network APs (physical topology information – distances from neighbors, network information – neighbors' IPs and MACs, algorithm information – neighbors' interference prices and TxPowers).
- The "CPC communication" software module consists of two parts, the client and the server. As mentioned afore, the basis of the CPC scheme is related to the asynchronous information exchange among the network elements. This implies that each network element operates as a server, where the neighboring WiFi APs are being associated and also as a client in order to associate to the neighboring APs.
- The "Control Engine" is responsible for the enforcement of the re-configuration action, which in the considered case is the TxPower adjustment.
- The "Monitoring" software module is responsible for the two types of monitoring tasks, the local and the neighborhood/cluster. The former is related to monitoring of local metrics and measurements (e.g., identification of local TxPower, associated users, sensed APs, etc.) whereas the latter is related to cluster information (e.g. MACs and IPs of neighboring APs, physical topology graph, etc.).

The afore-described software comprises the CPC application that has been deployed in every CPC-enabled network element.

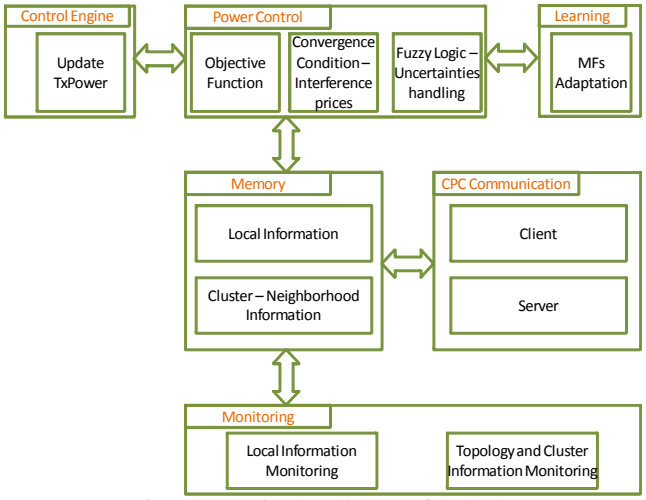


Figure 3: Functional architecture of the CPC

B. Case Study

In the case study under investigation, we assume the presence of several WiFi APs located in the considered area. These APs communicate via wireless links in order to exchange their interference values. Based on these values each network element adjusts its transmission power (Fig. 4).

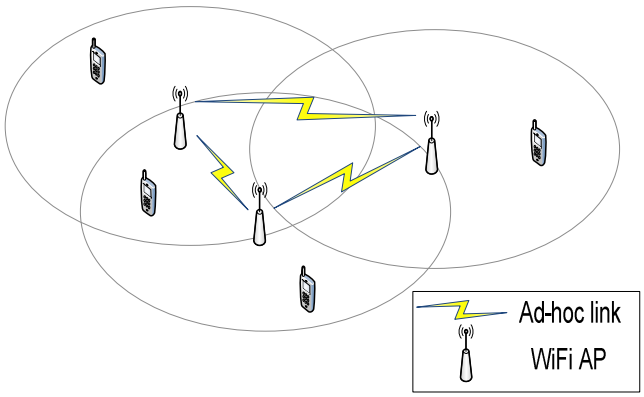


Figure 4: Envisaged network topology

Given the assumption that the APs communicate asynchronously and each one might have its locally-set update period, it is possible that the APs are unaware of the current network's status (from the messages exchange). Such problem becomes even more acute if we consider that the network elements might lose some messages during the messages exchange procedure due to the nature of the applied information fusion scheme and the sensitivity of the wireless medium. This implies that the use of the fuzzy reasoner is imperative for capturing the uncertainties [14]. The WiFi application area though, poses the need for modification of the inputs and the inference engine of the fuzzy logic controller. Thus, the number of the WiFi APs in the vicinity, the number of users in the vicinity (associated to WiFi APs) and the update interval are used as inputs of the fuzzy reasoner. In case of completely new application areas, new/modified fuzzy reasoners could be incorporated so as to be more suitable to the use case under discussion. The way a network element perceives its environment is based on the input and output membership functions. As in [14], the inputs' membership functions initially are set to have triangular shape, mainly in order to capture the strict nature of the inputs.

Table I provides the rules of the inference engine of the fuzzy reasoner. The most crucial input for the decision making process is the update interval. This input depicts the frequency of the information updates about the interference price of a network element to its neighbors thus capturing how recent is the view of a network element, based on the inputs from its neighbors. These inputs will be used for the calculation of the TxPower (Section IV).

TABLE I. RULES OF THE FUZZY REASONER

Rule Number	Num of WiFi Aps	Num of Users	Update Interval	Interference price
1	Low	Low	Low	Low
2	Low	Low	Medium	Low
3	Low	Low	High	Medium
4	Low	Medium	Low	Low
5	Low	Medium	Medium	Medium
6	Low	Medium	High	Medium
7	Low	High	Low	Medium
8	Low	High	Medium	Medium
9	Low	High	High	High
10	Medium	Low	Low	Low
11	Medium	Low	Medium	Medium
12	Medium	Low	High	High
13	Medium	Medium	Low	Medium
14	Medium	Medium	Medium	Medium
15	Medium	Medium	High	High
16	Medium	High	Low	Medium
17	Medium	High	Medium	Medium
18	Medium	High	High	High
19	High	Low	Low	Medium
20	High	Low	Medium	Medium
21	High	Low	High	High
22	High	Medium	Low	Medium
23	High	Medium	Medium	Medium
24	High	Medium	High	High
25	High	High	Low	Medium
26	High	High	Medium	High
27	High	High	High	High

As briefly described in Section IV, the CPC consists of two separate iterative procedures, the power update and the interference price update. In the former, consider a network element i, which updates its transmission power using a time interval $t_{ai} \in T_{ai}$, where T_{ai} is a set of positive time instances in which the AP i will update its transmission power level and $t_{a1} \neq t_{a2} \neq \cdots \neq t_{ai}$. Similarly, each WiFi AP *i* has an interference price update interval $t_{bi} \in T_{bi}$, where it updates its interference price and announces the updated interference price π_i^k to the rest of the WiFi APs belonging in the scheme. Fig. 5 provides the messages exchange and the operations' sequence on a scheme with two WiFi APs; this could be generalized for more APs as well.

C. *Learning Algorithm*

The proposed learning algorithm consists of three parts, namely, the monitoring/labeling, the classification and the adaptation of the fuzzy reasoner. Each network element that is part of the network monitors its own environment. Every time that the network elements collaboratively proceed in transmission power adjustment, their interference prices are being compared to the previous ones and the interference factor calculations are being labeled as:

- Beneficiaries: for the decisions that led to reduction of the interference value caused to the neighboring network elements,
- Neutral: for the decisions that led to similar interference values. In such cases the decision could not be characterized either as correct or wrong,
- Non Beneficiaries: the decision led to an increase of the interference value caused to the neighboring network elements.

More specifically, periodically, the network elements cooperatively identify the optimum transmission power using the methodology described in Section IV; the iterative procedure requires finite number of steps (i.e., maximum 30 iterations). Before every periodic transmission power adjustment, the interference value is being compared to the value before the last transmission power adjustment (Fig. 6).

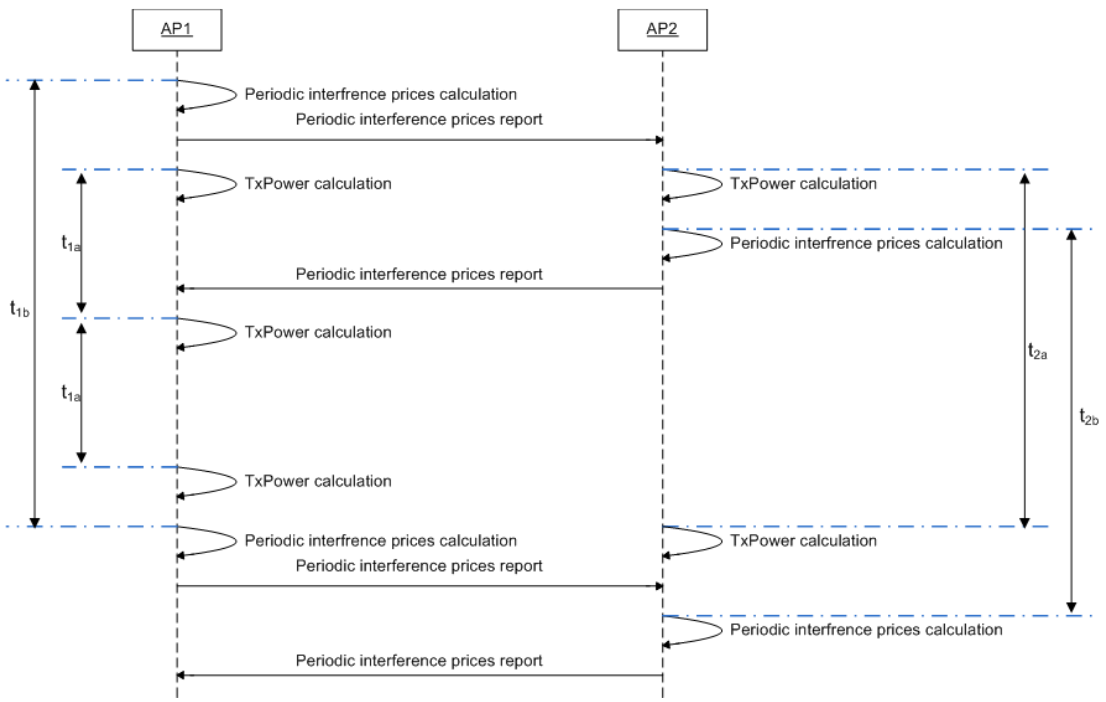


Figure 5: Message sequence chart for two WiFi APs

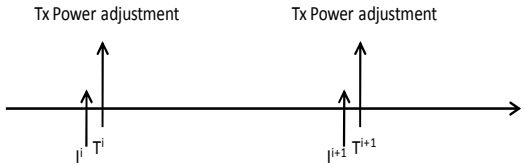


Figure 6: Timeline for Interference calculation and transmission per adjustment

The input vector $\vec{Z}_i$ (i.e., num of WiFi APs, num of users, update interval) of each network element is being evaluated against a predefined fuzzy inference system and results to an *a* value which, in conjunction to the interference prices, is used for the calculation of the optimum transmission power. Comparing the interference prices just before the initiation of the i^{th} transmission power adjustment and the $(i+1)^{th}$ we label the decision accordingly (i.e., Y_i is beneficiary, neutral or non beneficiary). The comparison is done using the Euclidian distance metric. This procedure results to a set (*S*) of labeled decisions which have been correctly labeled (at a great level of certainty) through the afore-described phase. Table II presents the key points of monitoring/labeling part of the developed algorithm.

TABLE II. MONITORING/LABELING ALGORITHM

Input:	**Approximation Parameter ε, Sample Size N**
Output:	**Set of observations S**
1.	$S \leftarrow O$
2.	$i=0$
3.	while true
4.1	i++
4.2	Retrieve vector $\vec{Z}_i$ and $\vec{IP}_i$
4.3	$\alpha_i \leftarrow$ fuzzy logic ({# WiFi APs, # Users, Update Interval})
4.4	Calculate TxPower
4.5	Wait for $\vec{Z}_{i+1}$ and $\vec{IP}_{i+1}$
4.6	Calculate I^{factor}_{i+1}
4.7	If ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert < \varepsilon$) $\rightarrow Y_i$=*Neutral* Else ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert > \varepsilon$) and ($I^{factor}_i - I^{factor}_{i+1} > 0$) $\rightarrow Y_i$ = *Beneficiary* Else ($\lvert I^{factor}_i - I^{factor}_{i+1}\rvert > \varepsilon$) and ($I^{factor}_i - I^{factor}_{i+1} < 0$) $\rightarrow Y_i$ = *Non Beneficiary*
4.8	$S \leftarrow S \cup \{ \vec{Z}_{i+1}, \vec{IP}_{i+1}, Y_i \}$
5.	return *S*

On sequence, we formulate three clusters using the labeled data in order to exclude the misclassfied data from the previous step; the clustering is performed using k-Means Thus, each network element maintains a set of three clusters, one for classifying every decision type. By representing each cluster to a 3D grid we map each cluster to a geometrical object (i.e., sphere S_i). Each sphere is centered at $C_j=\Sigma_{i=1}^{|Ci|} S_i/|C_i|$ and has radius $R_j=max_{i=1}^{|Ci|}||CE_i-S_i||$.

For each couple of clusters *i*, *j*, the cluster centers C_i, C_j define a line ε that interconnects the two points. This line can be described by the following equation:

$$p_m = x_m + u \cdot (y_m - x_m),\ m = 1...d \quad (7)$$

Line ε intersects with spheres S_i and S_j in four points which can be retrieved by substituting the p_m values into the following hypersphere equations:

$$D_i \rightarrow \sum_{m=1}^{d} (p_m - x_m)^2 = R_i^2 \quad (8)$$

$$D_j \rightarrow \sum_{m=1}^{d} (p_m - y_m)^2 = R_j^2 \quad (9)$$

A simple way of identifying the bounds would be to extract the intersection points which belong to different hyperspheres and exhibit minimum distance from each other [16]. Then, as shown in Fig. 7, we map the identified bounds to the input membership functions of the fuzzy reasoner; this results to the modification of the environment perception of each network element.

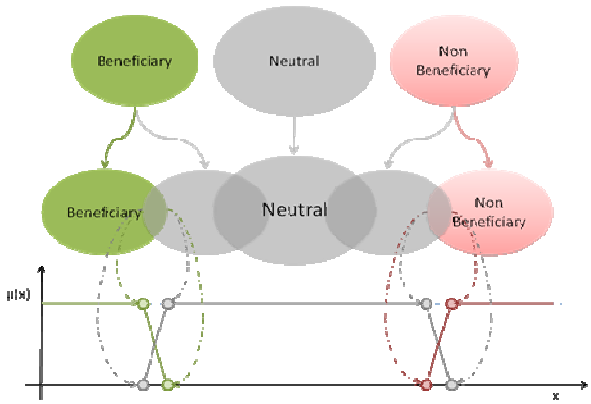


Figure 7: Clustering and bounds extraction mechanisms

VI. DEPLOYMENT

In order to experiment with the developed solution, we have proceeded in a series of real life experimentations in our premises. For this purpose we have used the proof of concept that we have implemented, which instantiates the algorithm described in Section V.

A. Environment Description

For the experimentation a set of Soekris devices has been used; such devices are low-power, low-cost, Linux-based communication computers (500MHz AMD Geode LX, 512MByte DDR-SDRAM) that act as re-programmable WiFi APs by using IEEE 802.11b/g radio access technology [17]. In all Soekris devices we consider two wireless interfaces, one is the actual AP interface and the other one is used for monitoring; the former is the AR5413 mini-PCI [18] Card whereas the latter is the WUSB54GC USB card [19]. The APs deploy their own network and route the information to the internet through NAT. APs are connected through the backbone network and communicate with a standalone machine which aggregates information and provided triggers for the initiation of algorithms. The CPC implementation is based on Java programming language using several external

libraries. The most important of them are the jFuzzyLogic [20] for the "Power Control" module and Apache MINA [21] for the "CPC communication" module. For the "Monitoring" module the Linux kernel utilities are exploited.

Four Soekris devices have been placed in our premises, which suggest a typical small office environment consisting of three rooms, with 15 researchers (Fig. 8(a)). The researchers used these APs for 3 consecutive days for 10 hours each day (from 10:00 CET until 20:00 CET on July 9th 2012, where our algorithms are not installed and the measurements are used for extracting the control data, and on July 10th and 11th 2012 where our algorithms operate for the transmission power control) in order to access the internet and perform all normal, working-day, activities. Overall traffic throughout the day ranged from 1 to 10 Mbps while APs were configured to operate at 5.5Mbps throughput. The network layout is depicted in Fig. 8(b).

In all three days of our experiment, the one for the control data generation and the two where the CPC was embedded in the Soekris devices, we have attempted to procedure almost identical experimental conditions. The bandwidth requirements were reproduced – however user's mobility could not be identically reproduced.

B. Assumptions

As mentioned afore, the CPC scheme is based on the assumption that it will operate on an urban area. Thus, the generic assumptions of the algorithm should be also adapted accordingly.

The WiFi APs are placed in an indoor environment and communicate via specific communication interfaces. This implies that the distance among the network elements needs to be defined. In the proposed approach, the methodology of [22] and [23] is being followed.

The propagation obeys to certain models, from which the log-distance model is one of the most simple; the following equation describes the behaviour of such model:

$$\log d = \frac{1}{10 \cdot n}(P_{TX} - P_{RX} + G_{TX} + G_{RX} - X_a + 20\log\lambda - 20\log(4\pi)) \quad (10)$$

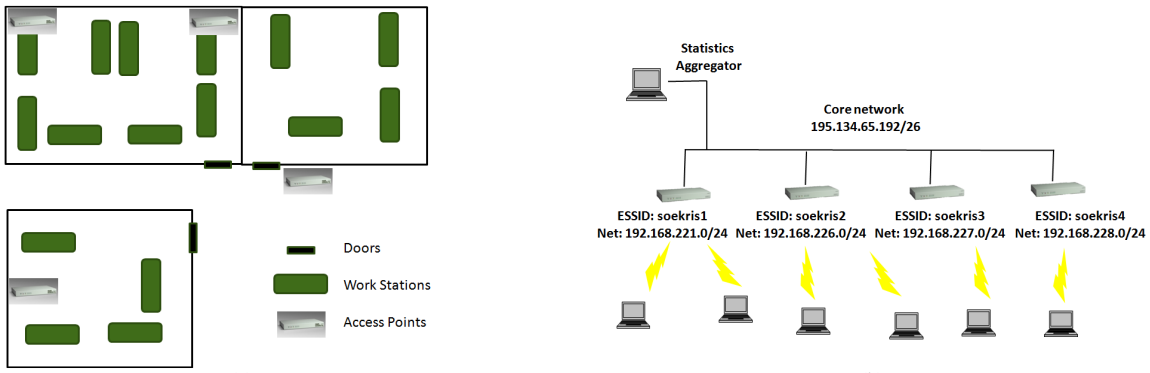
Where

- d (m): the estimated distance between the transmitter and the receiver,
- P_{TX} (dBm): the transmitted power level,
- P_{RX} (dBm) is the power level measured by the receiver,
- G_{TX} (dBi): the antenna gain of the transmitter,
- G_{RX} (dBi): the antenna gain of the receiver,
- n: measure of the influence of obstacles like partitions and ranges from 2-5 (2 for free space, 4-5 in case obstacles are considered),
- X_α: normal random variable with standard deviation of α. This variable captures the variance of the fading phenomena in an indoor environment,
- λ (m): the wavelength of the signal (for WiFi can be considered 0.12).

In the proposed experimentation, and for a typical office environment, n has been set to 5 and X_α to 20. Regarding the transmission power, which is the actual parameter of our implementation, it is related to the equipment's capabilities. Specifically, TxPower, is limited by the WiFi card's capabilities; 10dBm is the lowest price whereas 27dBm is the highest.

VII. Experimentation Analysis

For the evaluation of the CPC algorithm, we have followed an extensive experimentation scenario, in a real office environment in order to validate the applicability of the proposed solution and also to highlight the energy and network benefits from the incorporation of our algorithms. The experimentation analysis moves towards two directions, on the testing of the CPC and its applicability in the use case under consideration (i.e., a realistic WiFi office environment)



(a) (b)

Figure 8: (a) Physical topology of the experimentation environment, (b) network topology of the experimentation environment.

and on the evaluation of the learning/adaptation capabilities.

Fig. 9 and 10 capture our experimentation results for the first day, where the CPC algorithm operates without the learning part; this implies that the first day of experimentations the algorithm operates in order to gather data which will be used for the adaptation of the situation perception (i.e., first day of the experimentations). The experiment has started 10:00 CET and has finished 20:00 CET. The four Soekris APs have been placed in our testbed and we have been measuring for this period the transmission power of their WiFi cards; the transmission power ranges from 10 to 27 dBm. Fig. 9 presents the transmission power for the 10 hours of the experiment. In order to evaluate the operation of the network for several topologies, initially we have all four Soekris operating, whereas as the experiment proceeds we turn them off one by one and we leave only one operational. For each of the Soekris devices (and considering that the 10dBm is the basis of the TxPower for each AP) we see the actual gain compared to setting the transmission power to the maximum TxPower (i.e., 27 dBm). The energy gain at each of APs 1, 2, 3, and 4 is 12.51%, 10.75%, 33.33% and 21.23% respectively. Also, it is obvious that the more the APs, the more energy gains we have, due to the collaborative nature of the algorithm. Also, what should be noticed is the fact that the APs change very often their TxPower levels. This is related to the highly volatile office environment, with moving users and the many interference sources (i.e., moving users, cell phones, Bluetooth devices, etc.), in relation to the fact that the APs identify the network topology considering indoor path loss models. Such models, if we assume static environments, without moving users operate with accuracy, however in the case under discussion, the network elements need to calculate the topology on a constant basis, in every CPC loop.

Fig.10 provides the 6th degree polynomial function of the SINR measurements during the experimentation. At any case, the SINR is better compared to the case where maximum TxPower has been set to the APs. For the AP 3 and 4 the experiment stops at the time that these APs are being turned off (13.20 and 14:20 respectively) and we see that when all four Soekris operate, the SINR to all of them is low. When we start turning off AP we observe that the SINR to all the operating ones starts increasing; this is related to the fact that the interference that is caused reduces as well. Finally, only one, AP 2, remains operational and we have a huge increase in the SINR, which has started when we turned off AP3 and AP4; however we should take under consideration that the overall capacity reduces.

Fig.11 presents the number of iterations every time the CPC is being triggered. We consider that the CPC is being triggered periodically, every 5 minutes. The Soekris APs exchange messages asynchronously; everyone using its own intervals. We observe that the scheme converges in small number of iterations most of the times (mean value of iterations 3.876).

The initial configuration of the network elements is a

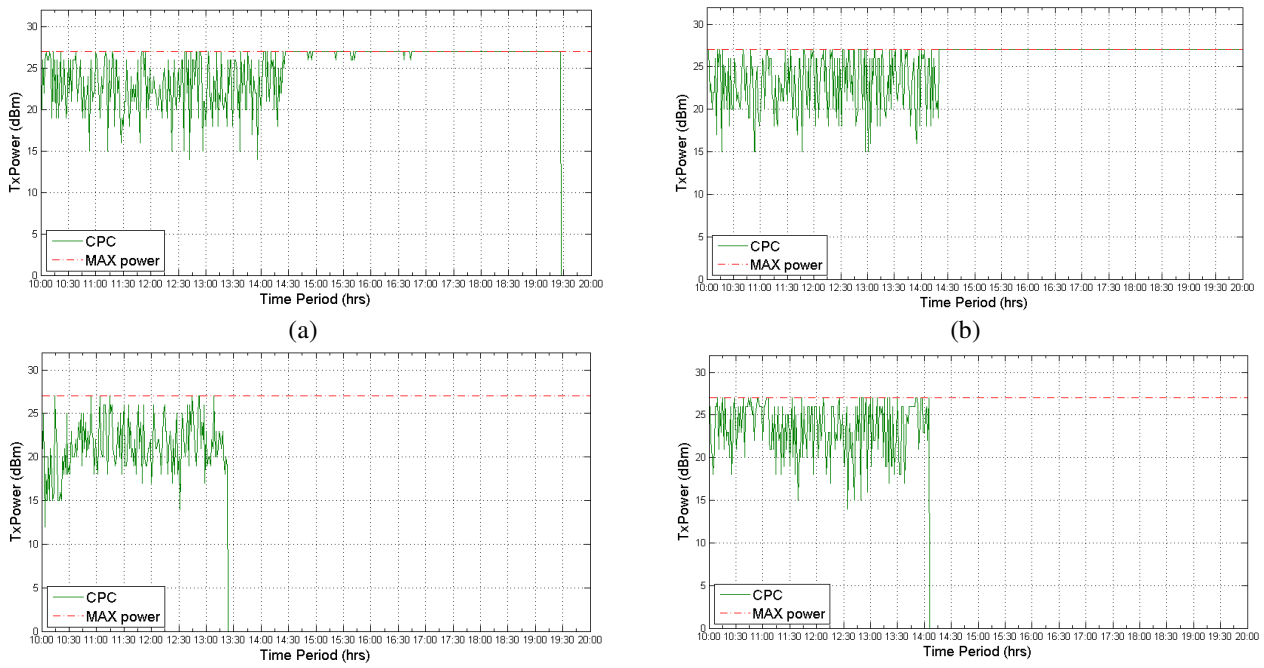


Figure 9: Transmission power adjustments in the four Soekris APs using the Cooperative Power Control scheme in (a) Soekris AP1, (b) Soekris AP2, (c) Soekris AP3, (d) Soekris AP4

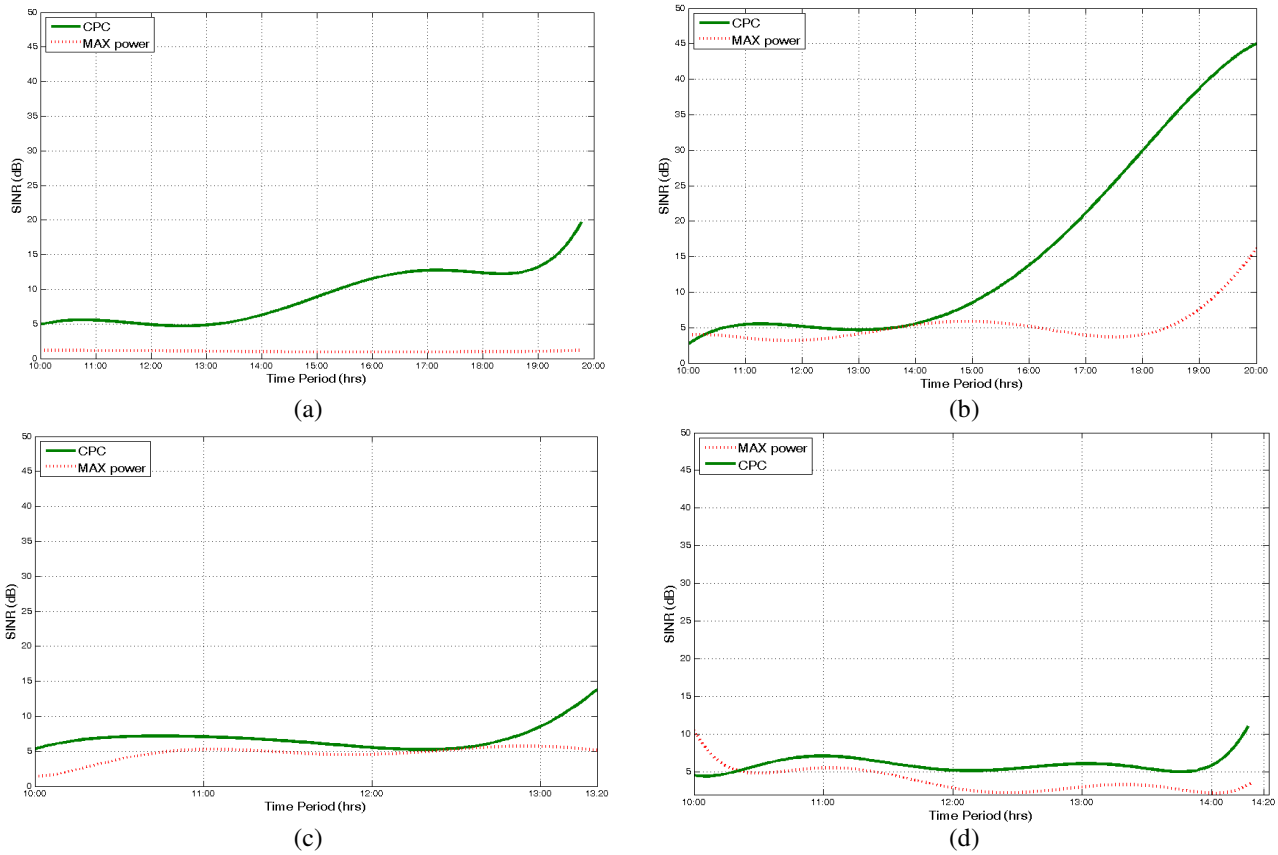


Figure 10: SINR evolution during the experimentation period for (a) Soekris AP1, (b)Soekris AP2, (c)Soekris AP3, (d)Soekris AP4

generic one and captures a great variety of environments. However, for both the physical and network topology which has been used for experimentation, this configuration is not the most suitable one. Thus, the adaptation scheme that has been presented in Section V.C has been incorporated.

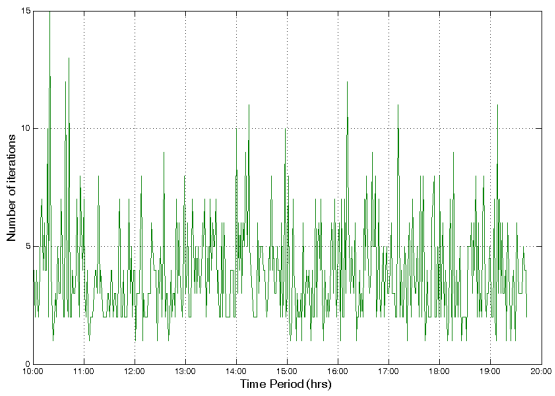


Figure 11: Number of iterations every time the CPC is being triggered

During the first experimentation day of the CPC in the Soekris devices, the inputs of the fuzzy reasoner are being collected. Then, the tuples are being clustered and the overlapping areas are being mapped to the uncertainty bounds in the input membership functions. Fig. 12 provides the transmission power throughout the second experimentation day for all Soekris devices, with the adapted input membership functions (learning-based CPC scheme).

As it is obvious, the CPC scheme is more sensitive to the environment, compared to the previous day of experimentations. Given the fact that they operate in the same environment, the APs proceed even more often in transmission power adjustments. Also, when only two APs remain operational, as the experimentation proceeds, we observe that they proceed in transmission power adjustments, according to the environment stimuli, contrary to the first day, where the transmission power adjustment mainly occurred when all the APs were operational.

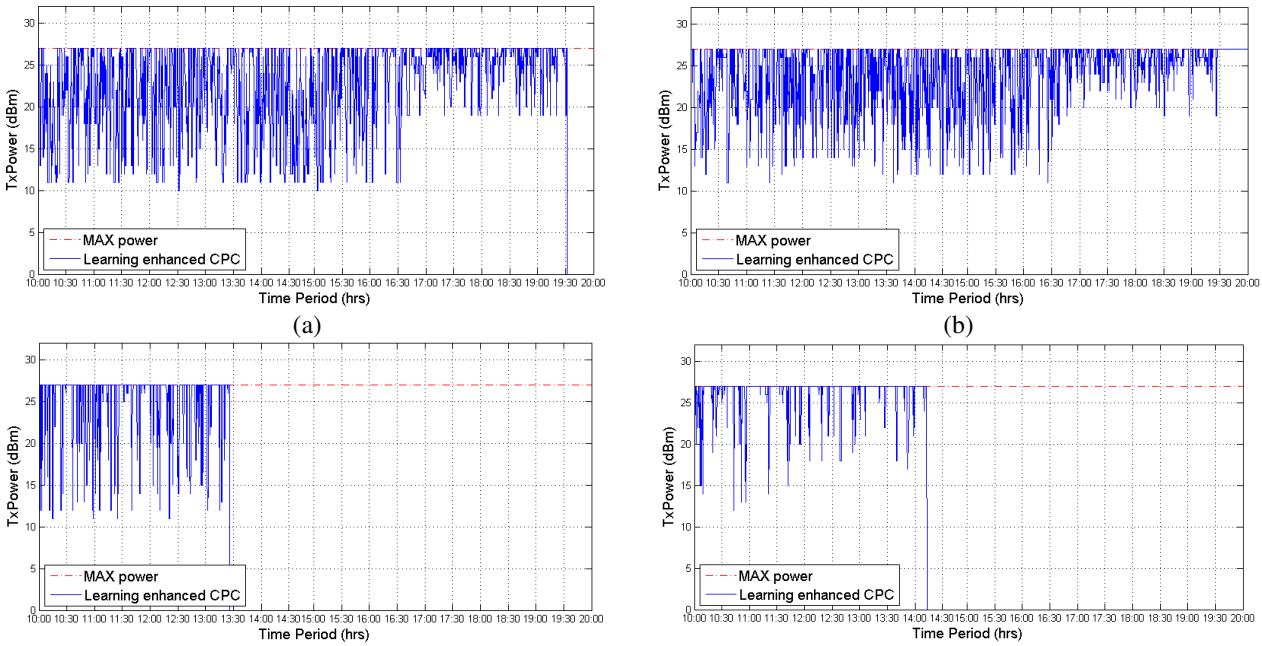


Figure 12: Transmission power adjustments in the four Soekris APs using the Learning enhanced Cooperative Power Control scheme in (a) Soekris AP1, (b)Soekris AP2, (c)Soekris AP3, (d)Soekris AP4

Furthermore, we observe significant energy gains, in relation to the case without learning capabilities. More specifically, AP 1 has 24.73% less power consumption compared to the maximum transmission power, whereas AP 2 consumes 18.01% less power, AP 3 14.69% and AP 4 5.65%. Given the fact that AP1 and 2 are the APs that remain operational almost throughout the experiment, we conclude that the energy gains are even more significant. Regarding the SINR, it remains in the same levels as in the case of the core CPC algorithm (Fig. 13), due to the fact that the objective function to be optimized is the same. The APs proceed in power adjustments in lower transmission power

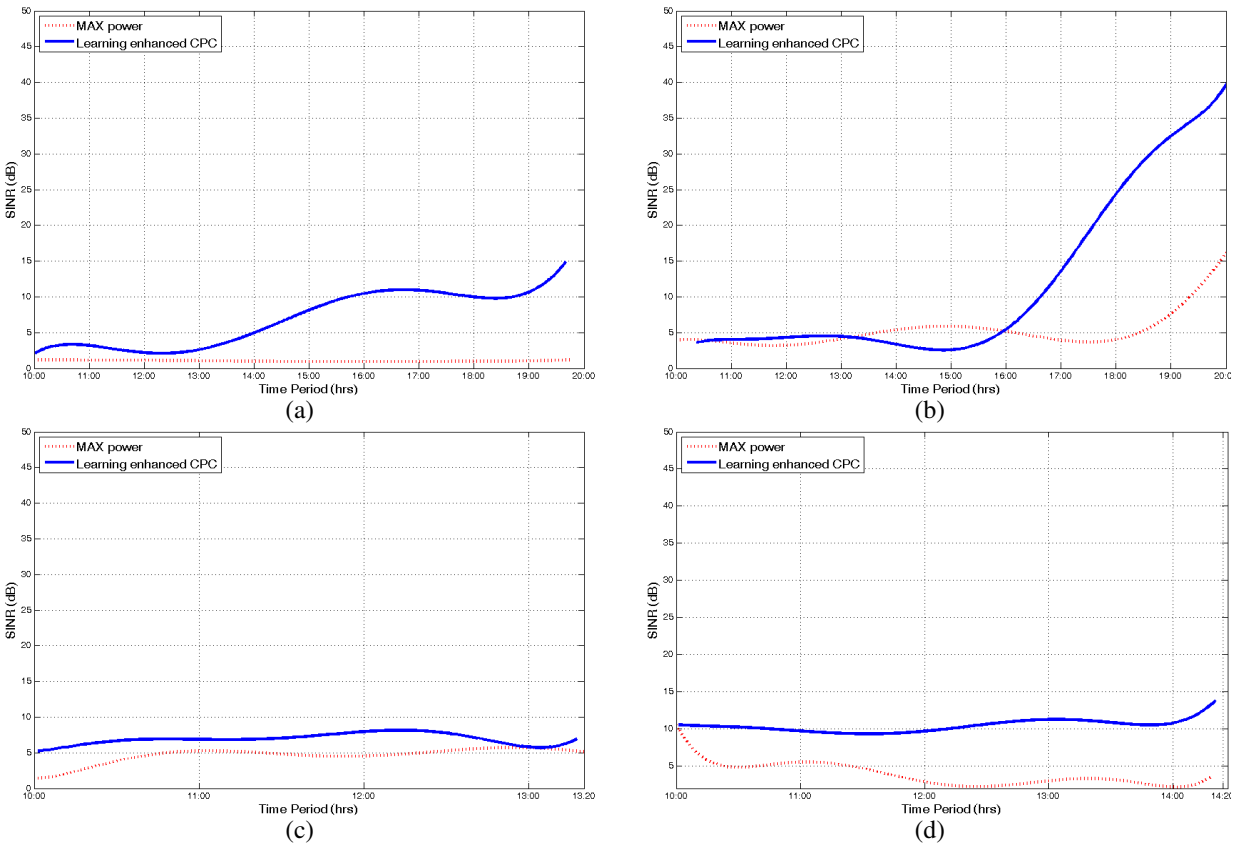


Figure 13: SINR evolution during the experimentation period for (a) Soekris AP1, (b)Soekris AP2, (c)Soekris AP3, (d)Soekris AP4

levels resulting in less interference as well; however the SINR remains at the same levels, due to the decrease in both metrics (i.e., TxPower and interference). Fig. 14 presents the number of iterations every time the CPC is being triggered after the learning procedure. Similarly to the core CPC, we observe that the scheme converges in small number of iterations most of the times; furthermore, we observe a slight decrease in the overall mean value of iterations (3.47) which also highlights that the system has become more suitable to its environment. Finally, considering that the algorithm is being triggered periodically, every 5 minutes for 10 hours, we observe that the adaptation algorithm enhances the situation perception scheme using relatively small amount of measurements (4 AP * 120 measurements/AP = 480 measurements).

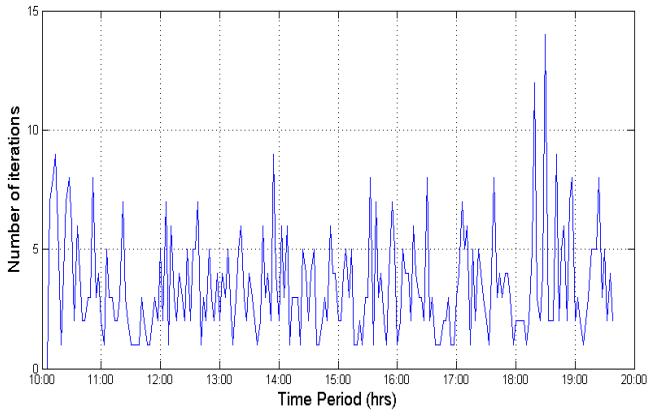


Figure 14: Number of iterations every time the CPC is being triggered

VIII. Conclusion And FutureWork

This paper proposes an algorithm for power control and interference mitigation. The solution also incorporates learning capabilities in order to enable the network elements to adapt to their situation perception according to the environment stimuli. The learning procedure captures the positive or the negative impact of an action (i.e., transmission power set value) in the interference that a network element causes to its neighbors.

The novelty of our contribution is the combination of the merits of fuzzy logic and data clustering for the optimal interpretation of the network uncertainties and its incorporation to the CPC algorithm. The network uncertainties have been identified using the cluster overlaps; the latter are then being translated in the environment perception of the fuzzy reasoners (i.e., input membership functions).

The proposed solution has been tested in a realistic office environment in a real life experiment. The algorithm has been deployed in WiFi APs and used for their transmission power control. The experimental analysis proved the applicability of the CPC and the benefits from its incorporation. More specifically, the network elements have significant energy gains by incorporating the CPC; the addition of the learning capabilities in the APs makes them more sensitive in the environment stimuli. The experimental analysis proved that also the WiFi APs SINR benefits from the incorporation of the CPC; in every case the SINR is improved compared to an environment where all APs set their TxPower to maximum levels. Thus, the network elements achieve higher SINR levels and also have energy gains by setting their TxPower to the most suitable level for them and for the overall network.

Our future work includes the incorporation of more sophisticated data mining techniques (e.g., Support Vector Machines, C-Means, Subtractive clustering, etc.) in order to have better adaptation to the environment. Furthermore, the incorporation of outlier detection techniques will be investigated in order to ensure that only valid measurements will be used for the learning procedure.

Acknowledgment

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement CONSERN n° 257542.

The authors would like to thank Dr. Panagis Magdalinos for his help in order to improve the paper and the SCAN Lab members for participating in the experimentation procedure.

References

[1] P. Spapis, G. Katsikas, M. Stamatelatos, K. Chatzikokolakis, R. Arapoglou, and N. Alonistioti “Learning Enhanced Environment Perception for Cooperative Power Control”, Fifth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2011), 2011.

[2] Q. Sun, X. Zeng, N. Chen, Z. Ke, and R. Ur Rasool, “A Non-cooperative Power Control Algorithm for Wireless Ad Hoc and Sensor Networks”, Second International Conference on Genetic and Evolutionary Computing (WGEC), 2008.

[3] J.F. Nash, "Equilibrium points in n-person games", Proceedings of the National Academy of Sciences 36(1): 48-49, 1950.

[4] P. C. E. Shannon, “Communication in the presence of noise,” Proceedings of the Institute of Radio Engineers, vol. 37, pp. 10–21, 1949.

[5] G. J. Foschini and Z. Miljanic, “A simple distributed autonomous power control algorithm and its convergence,” IEEE Trans. Veh. Technol., vol. 42, pp. 641–646, 1993.

[6] D. Mitra, “An asynchronous distributed algorithm for power control in cellular radio systems,” in Proc. 4th Winlab Workshop Third Generation Wireless Information Network, pp. 249–257, 1993.

[7] Chun-Gang Yang, Jian-Dong Li, and Zhi Tian, “Optimal Power Control for Cognitive Radio Networks Under Coupled Interference Constraints: A Cooperative Game-Theoretic Perspective”, IEEE transactions on vehicular technology, vol. 59, no. 4, pp. 1696-1706, 2010.

[8] V.P. Mhatre, K. Papagiannaki, F. Baccelli,"Interference Mitigation Through Power Control in High Density 802.11 WLANs", 26th IEEE International Conference on Computer Communications. (IEEE INFOCOM 2007), pp.535-543, 2007.

[9] K. Ramachandran, R. Kokku, Honghai Zhang, and M. Gruteser, "Symphony: Synchronous Two-Phase Rate and Power Control in

802.11 WLANs," IEEE/ACM Transactions on Networking, vol.18, no.4, pp.1289-1302, Aug. 2010.

[10] K. Kowalik, M. Bykowski, B. Keegan, and M. Davis, "An evaluation of a conservative transmit power control mechanism on an indoor 802.11 wireless mesh testbed," International Conference on Wireless Information Networks and Systems, (WINSYS'08), 2008.

[11] Y. Wei, M. Song, and J. Song, "An AODV-improved routing based on power control in WiFi mesh networks," Canadian Conference Electrical and Computer Engineering 2008, (CCECE 2008), pp.001349-001352, 2008

[12] T.A. ElBatt, S.V. Krishnamurthy, D. Connors, S. Dao, "Power management for throughput enhancement in wireless ad-hoc networks," International Conference on Communications, (IEEE ICC 2000), pp.1506-1513, 2000.

[13] J. Huang, R. Berry, and M. Honig,"Spectrum sharing with distributed interference compensation", First IEEE International Symposium New Frontiers in Dynamic Spectrum Access Networks (DySPAN), 2005.

[14] A. Merentitis and D. Triantafyllopoulou,"Transmission Power Regulation in Cooperative Cognitive Radio Systems Under Uncertainties", IEEE International Symposium on Wireless Pervasive Computing (ISWPC), 2010.

[15] J. Han. and M. Kamber, "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management System. ISBN-13: 978-0123814791

[16] P. Magdalinos, A. Kousaridas, P. Spapis, G. Katsikas, and N. Alonistioti, "Feedback-based Learning for Self-Managed Network Elements", 12th IEEE International Symposium on Integrated Network Management, (IM2011), 2011.

[17] http://soekris.com/products/net5501.html, Dec. 2012.

[18] http://homesupport.cisco.com/en-eu/support/adapters/WUSB54GC, Dec. 2012.

[19] http://www.pcengines.ch/pdf/wlm54ag23.pdf, Dec. 2012.

[20] http://jfuzzylogic.sourceforge.net/html/index.html, Dec. 2012.

[21] http://mina.apache.org/, Dec. 2012.

[22] A. Bose and F. H. Chuan, "A practical path loss model for indoor WiFi positioning enhancement," 6th International Conference on Information, Communications & Signal Processing, 2007, pp.1-5, 2007.

[23] J. S. Seybold, "Introduction to RF propagation", Wiley, ISBN-13 978-0-471-65596-1.

An Interoperability Service for Autonomic Systems

Richard John Anthony
The University of Greenwich
Park Row, Greenwich
London SE10 9LS, UK
+44 (0) 208 331 8482
R.J.Anthony@gre.ac.uk

Mariusz Pelc
The University of Greenwich
Park Row, Greenwich
London SE10 9LS, UK
+44 (0) 208 331 8588
M.Pelc@gre.ac.uk

Haffiz Shuaib
The University of Greenwich
Park Row, Greenwich
London SE10 9LS, UK
+44 (0) 208 331 8588
Haffiz.Shuaib@yahoo.com

***Abstract* - Interoperability support is a key outstanding requirement for autonomic computing systems, and this need stems from the very success of these systems. Autonomic computing is increasingly popular; soon autonomic control components will be commonplace and present in almost every large or complex application. Interoperability between autonomic managers is an increasingly urgent concern, as the proliferation of autonomic systems inevitably leads to situations where multiple autonomic components coexist and interact either directly or indirectly within the same application or system. Problems can arise when numerous *independently* designed autonomic components interact, potentially destabilising systems. We advocate a service-based approach to interoperability and present a set of requirements for such an approach as well as a suitable architecture. A key component of this architecture is the Interoperability Service with which Autonomic Managers register their management interests and capabilities, using a management description language. The Interoperability Service automatically discovers and manages potential conflicts between manager components. Developers integrate Autonomic Managers with the Interoperability Service by importing its interfaces. This allows the Interoperability Service to automatically suspend and resume managers, or specific management functions as necessary, driven by the automated conflict detection. We illustrate the use of the Interoperability Service in a data-centre scenario in which independently developed power management and performance management autonomic components operate.**

Keywords - Autonomic systems; Interoperability; Services.

I. INTRODUCTION

Autonomic Computing (AC) has matured rapidly from a hot research topic to an accepted and valued technique for automating system management, in less than a decade. The main reason that the popularity of AC has grown so strongly in such a short timeframe is because it offers solutions to the problems caused by high complexity in systems. This complexity arises from large numbers of interacting components, typically with high functionality and with high operational speeds working in high throughput applications. The number of possible configurations and the different interactions and sequences of interactions, increases at an exponential combinatorial rate as the underlying behavioural richness of the systems and sub-components increases. This rapidly leads to systems whose behaviour is beyond a human manager's comprehension, certainly in terms of making real-time configuration decisions. Autonomic computing automates the management of one or more sub-components or resources, thus controlling certain elected characteristics of a system in a timely manner; increasing optimality and robustness and reducing errors. The sophistication of AC has also advanced at a spectacular rate. This is largely due to the reuse and extension of a wide range of reasoning and control concepts and techniques taken from established fields such as control theory and artificial intelligence.

The rapid evolution of AC has been driven by a main focus on the internal reasoning techniques, and a bias towards isolated development and deployment of Autonomic Managers (AM) which tend to have a very specific operational envelope; in order to demonstrate the robustness of the core techniques and thus to gain acceptance for the overall concept of AC.

However, the popularity of AC is driving expansion into ever more diverse application domains and increasing the variety of aspects of systems that can be automatically managed. This means that for future AMs, it is not safe to assume isolated management operation. In fact, it will be increasingly common for multiple AMs to coexist in any moderately sized computer system.

Almost all systems use multi-vendor software solutions and this implies that there will be potentially a variety of manager components existing, even for any one specific function of a system. For many systems, autonomic management will arrive incrementally; as new functionality is introduced, and through upgrades of non-managed components to new managed versions. In some cases the introduction of management capabilities will not be obvious – third party developers may deliver components with internal management that is not exposed at interfaces to other components.

Unplanned coexistence, or unexpected interactions could arise due to the highly dynamic nature of some systems in which configurations, and composition of components changes quickly. Automatic upgrades of individual components are another increasingly popular way by which systems behaviour changes over time, and not necessarily with the designer of a specific component having full

visibility of the whole system behaviour. Thus even a 'known' manager component could suddenly introduce new behaviour or potential conflict.

The possibility of coexistence and thus unplanned interactions or resource conflicts means that AMs will operate in environmental conditions not foreseeable by their designers. This means that an AM may pass behaviour tests 'in the lab' but still exhibit undesired behaviour when deployed.

This work extends our earlier work in [1]. We are interested in the challenge of interoperability for AMs, especially in the context of unplanned interactions, which can take many forms, but fall into two classes. *Direct* conflicts occur where two AMs attempt to manage the same explicit resource. *Indirect* conflicts arise when AMs control different resources, but the management effects of one have an impact on the management function of the other, or the combined effect of the two managers has an undesirable impact at system level.

The indirect conflicts are expected to be the most frequent and problematic, as there are such a wide variety of unpredictable ways in which such conflicts can occur. In addition, the effects of indirect conflict will be less obvious to detect and harder to diagnose than the direct conflicts. There will also be a range of severity of the effects of conflicts, from little consequence (such as a cancellation effect of opposing managers) whilst others could lead to serious performance or stability problems or even failure. The problem is illustrated with an example: Consider a system with two AMs: a Power Manager (PM1) shuts down servers that have been idle for a short time; and a Performance Manager (PM2) attempts to maintain a pool of idle servers to ensure high responsiveness to high priority applications. The two services were developed and evaluated in isolation and both performed perfectly; however the respective vendors did not envisage that they would co-exist. In current state of practice for AM development, interoperability is not a first-class concern, so each manager will be unaware of the other, i.e., it has no mechanism to detect and adapt to the presence and behaviour of the other. Bringing a shutdown server back on line has a latency of several seconds, thus when both AMs are co-resident PM1's 'locally correct' behaviour defeats PM2's contribution.

This problem can only be resolved if an external agent (such as a human system manager) can detect, diagnose, and identify a solution to the problem. This illustration is quite similar to the situation described in [2], see section II.

The general lack of interoperability support for AC is an urgent problem that could threaten its long-term success if not addressed in the near future. Custom solutions for interoperability may be necessary in some specific applications but in general this is a very expensive approach. In addition to the application-technical challenge, the interoperability solution itself becomes an additional component to keep up to date, as the AMs themselves, and the operating environment change over time. Some important issues arising from custom interoperability attempts are discussed in section II.

We advocate a universal solution for AM interoperability that is integrated into AMs at design time but which does not impose any limitations on the technology used to implement the management control functions and does not restrict or interfere with the way in which the autonomic management logic operates. We propose an Interoperability Service (IS) that monitors the various autonomic components present in a system. When a conflict of interest is detected the IS selectively suspends or shuts down the management function of autonomic components, based on a service description exchanged during the AM registration process (i.e., at run time). The IS has a hierarchical structure to ensure scalability and operates with a primarily local focus but also handles conflicts between non-local components where relevant. The proposed approach requires that at design time the developer identifies the resources that the manager will directly control, as well as those that could be indirectly affected. The approach has the main benefit of not requiring the developer to have any knowledge of other managers that may be present at run time. Compliance with such a scheme will be a step towards eventual 'certification' of AMs, which is important for long-term acceptance and growth of AC.

The contributions of this paper include: firstly we evaluate the nature and scope of the interoperability challenge for autonomic systems and identify a set of requirements for a universal solution (section III). We present the architecture of a service-based interoperability solution in section IV. Section IV, part C outlines a management description language which is intended for use by developers to ensure consistent description of AMs' management capabilities. Automatic detection of management conflicts is discussed in section IV, part D. Section V presents a work-in-progress implementation of the IS, and this is evaluated in section VI.

II. BACKGROUND

This section discusses the state-of-the-art in autonomic component interoperability. We also discuss some scenarios reported in the autonomic computing literature where either: purposeful interaction between several autonomic elements has been attempted to achieve a common goal; or where unexpected interactions or conflicts occurred between independent autonomic elements.

The potential significance of unwanted interaction between multiple autonomic elements was demonstrated in [2]. In this work, two autonomic managers were implemented. The first of these managers, the WebSphere Extended Deployment (WXD) dealt with application resource management, specifically in the area of CPU usage optimization. The second manager referred to as the Power

manager was responsible for modulating the operating frequency of the CPU to ensure that the power cap was not exceeded. It was shown that without a means to interact, both managers throttled and sped up the CPU without recourse to one another, thereby failing to achieve the said optimization the managers were expected to achieve, in terms of resource allocation and power utilization optimization, and potentially destabilising the system. We envisage widespread repetition of this problem until a universal approach to interoperability is implemented.

There are several examples of bespoke interoperability solutions for specific systems. A distributed management framework that seeks to achieve system-wide Quality of Service (QoS) goals for autonomic/self-managing systems was proposed in [3]. In this work, autonomic controllers were added and removed from the system based on the demands of the application QoS requirements. Here, the controllers communicate indirectly with one another using the system variables repository. If a controller were to fail, other controllers reading this repository take over the responsibilities of the failed controller, to ensure that QoS objectives are met. Other research works take a more direct approach to autonomic element interaction. For instance, in [4] the autonomic elements that enable the proposed data grid management system communicate directly with one another to ensure that management obligations are met. This paper defines four types of autonomic element including a data scheduler, data replication service provider, client and server file system providers. The relationship between each type of autonomic element is peer-to-peer. In contrast, [5] adopts a three-level hierarchical relationship to autonomic element interactions. The hierarchy is such that it is made up of a single device at its lowest level. Multiple devices are grouped into servers and servers are further grouped into clusters. The autonomic element at each level interacts with the autonomic elements above and below it to achieve autonomic power and performance management. [6] proposes a two-level autonomic data management system that optimizes the managed system so that jobs are not starved of resources. Physical servers each support multiple virtual servers. Local autonomic controllers manage each virtual server. These controllers use fuzzy logic rules to determine the expected amount of resources needed by the applications that run on the virtual servers. A global manager is tasked with allocation of physical resources to the virtual servers in an optimal and equitable manner. [7] implements a mechanism similar to that proposed in [6], in that virtualization on each physical server is used to optimize system usage and power consumption. The difference is that in [7] the local controllers manage each physical server as opposed to the virtual machine (VM) in [6]. A higher-level autonomic manager interacts with the local controllers to switch on or off the physical servers to ensure that Service Level Agreements (SLAs) are met, while also lowering power consumption. In [8] a combination of database replication and the avoidance of 'hot-spots' (devices with above-average operating temperature) is used to improve the performance of the managed system. Here, the autonomic system consists of two types of element. The responsibility of the first autonomic element i.e., the application scheduler is the creation and destruction of replicas of a database to assure high-availability. The other autonomic element, the resource manager, interacts with the scheduler to provide physical computational resources to the applications based on the SLAs. In addition to other responsibilities, the resource manager uses a model of past operations to move jobs from equipment operating at a higher temperature onto equipment with lower operating temperature. [9] describes an experiment to separate out the Monitoring and Analysis stages of the MAPE loop into distinct autonomic elements, with designed-in interactions between them. Monitoring capabilities are implemented in a node called an agent, with the analysis aspect implemented in a node called a broker. Information received from the environment are processed by the agents and forwarded to the broker where it is further analyzed. One or more agents feed information to a specific broker. An example of bespoke designed-in interaction between autonomic elements is provided in [10]. Three types of autonomic elements work hierarchically to provide scalable management, differentiated in terms of their operating timescale and scope of responsibility. This example serves to differentiate interaction between components which is achieved here, from the concept of interoperability which has stricter requirements. The fact that the various elements are part of a single coherent service with designed-in support for interaction means that the full challenge of interoperability is not encountered in this situation.

[11] illustrates the complexity of combining multiple management domains into a single controller. In this work a joint QoS and Energy manager is developed using a design-time oriented approach tuned for a specific environment and is thus highly sensitive to its operating conditions. This tight integration approach is not generalisable and the resulting combined manager would appear to be much more costly to develop and test than two independent managers.

The majority of the work to date has targeted planned interoperability between designed-for-collaboration AMs working towards a common goal. This is a valuable step towards AM interoperability, although these solutions generally lack a formal definition of the interfaces or where defined, these interfaces are highly specific to the system in question, thus preventing wide applicability and reusability.

Custom solutions are expensive to develop and are sensitive to changes in the target systems, thus they are generally restrictive and not future proof. A significant issue is that they do not tackle the specific problem of unintended or unexpected interactions that can occur when independently developed AMs co-exist in a system. However, the wider problem of standardised and system independent interoperability in autonomic systems has been

considered in several works. For instance, [12] defines a number of interfaces {Monitoring and test, Lifecycle, Policy, Negotiation and binding} to aid autonomic element interactions. Together these interface definitions enable the following properties:

- A means to establish appropriate administrative relationships.
- A means to monitor an autonomic element.
- A means to instruct these elements from an external source.
- A means to determine the current state of an autonomic element e.g., start, stop etc.
- A means to export and import policies to and from an autonomic element.
- A means to grant and request service to and from another autonomic element.
- A means to provide interaction integrity.

Multi-agent systems have some similarities to multiple independent-AM systems. However the interoperability problem is different because a multi-agent system is usually a coherent application and thus designed and tested specifically with the intention of multiple, similar, known-at-design-time agents; whereas in the independent-AM case incremental addition of new or upgraded AMs introduces unplanned interactions (i.e., unplanned at the time the various AMs were designed and tested).

Several 'vision' papers [13], [14], [15] identify interoperability as a key challenge for future autonomic systems. [13] argues that the mechanisms that define interoperability between autonomic elements must be reusable to limit complexities i.e., it must be generic enough to capture all communications across the board but also prevent bloatedeness. A standard means must exist for exchanging contexts between communicating elements to allow one autonomic element to understand the basis for the action of another autonomic element. [13] also identifies the need for a function to translate the output of one element to the format understood by another. [14] identifies some necessary components for autonomic element interaction, including: a name service registry for autonomic elements; a system interaction broker and a negotiator. An interface specification must also take cognizance of hierarchy amongst autonomic elements. [15] observes that a strict and specified communication behaviour should be enforced, to prevent interoperating autonomic elements from communicating through undocumented or backdoor interfaces.

III. INTEROPERABILITY ISSUES AND REQUIREMENTS

This section highlights the technical challenges of providing interoperability between AMs, and analyses the requirements for a universal solution. The state-of-the-art in achieving interoperability in autonomic systems has been discussed in section II and is predominantly focussed on custom and system-specific (or application-specific) solutions. This demonstrates the plausibility of AM interoperability and provides important starting points towards our goal of universal interoperability.

We posit that interoperability support (or lack of it) will become a make-or-break issue for future autonomic systems which inevitably contain multiple AM components. Bespoke or application-specific approaches to interoperability only offer a temporary respite at best, as they suffer a number of significant limitations which include:

1. Lack of flexibility and ability to scale - it is unrealistic to keep adding signals and functionality to deal with each possible interaction between any combination of AM's.

2. Having many isolated pools of interoperability is too complex. AC became popular fundamentally as a means of controlling, or hiding, complexity. It is undesirable from maintainability and stability perspectives to actually add excessive complexity in the process of solving the complexity problem.

3. It is not technically feasible to achieve close-coupled interoperability (i.e., where specific actions in one AM react to, or complement those of another) unless the source code and detailed functional specification is available for each AM involved. Without standardised interfaces this will always be a major challenge.

4. It will not be cost effective or timely. The cost and complexity of a bespoke solution spirals exponentially as the number of interacting AM's increase (consider a cloud computing facility or data centre with multi-vendor management software systems and with autonomic management embedded into platforms, operating software, application software and also infrastructure such as power management and cooling systems – this is a complexity and stability storm just waiting to happen).

5. Re-development of managers to facilitate specific interoperability, and especially to deal with conflicts that arise unexpectedly, is reactive and incremental (thus always ongoing).

6. It is not possible to know the nature of AMs not yet built, or to predict exactly if/where/when conflict will materialise in advance of adding a particular AM into a running system.

7. The incremental re-development approach cannot be applied on-line (in the medium term) as current technology is not sufficiently sophisticated, although for the longer term it may be possible since work is underway in several projects to develop self-evolvable systems.

In summary, the biggest single challenge to universal interoperability of autonomic systems is that it is not possible (at time of design, development or deployment of a particular AM) to predict all future autonomic services that could be added to a particular system, or even to predict upgrades that could be made to known services.

A. Requirements of a Universal IS

The issues highlighted above strongly suggest that it is necessary to deal with interoperability proactively by developing managers that are interoperability-enabled from the outset. We propose a service-based approach to interoperability, in which an Interoperability Service (IS) is responsible for detecting possible conflicts of management interest, and granting or withholding management rights to specific AMs as appropriate. In this way the IS performs all of the active interoperability management, and AMs only participate passively by providing information and following control commands from the IS. The IS interacts with AMs via a special interface which they must support.

We identify a number of requirements for a universal IS solution:

- Be application-domain independent and system independent.
- Able to represent AMs' management interests in a standard way that facilitates accurate conflict detection. This includes recognising resources which are not directly managed, but are nevertheless impacted by the behaviour of the manager.
- Have variable conflict-detection sensitivity which is run-time configurable to suit specific system requirements.
- Have a hierarchical architecture so as to deal with both local and global conflicts, and conflicts that occur across different levels in a complex system.
- Be proactive and automated; these are mandatory qualities for sustainable systems containing dynamic combinations of AM's with potentially complex interaction patterns.
- Able to automatically suspend and resume AM management activity on the basis of conflict detection and resolution.
- Support independently developed and tested AMs which in the presence of other AMs are susceptible to conflicts that they cannot locally detect or handle.
- Be sufficiently trustworthy that compliant AM's are *certifiable* for safe co-existence – regardless of platform, vendor etc.

Two diverse candidate architectural approaches were considered: The first is fully distributed, with localised conflict detection logic embedded in each autonomic manager. This approach requires that each manager exchanges standardised management description information with other managers on a peer-peer basis. Each participant would compare their own management interests with those of its discovered peers. On discovery of a conflict, a negotiation phase would determine which manager has the authority to manage the contested resource. This approach has the benefit of a standardised conflict detection mechanism, embedded in the form of a library, but has the disadvantages of extensive replication of functionality, the need for the negotiation phase, and potential scalability limitations.

The second approach is central service based. This approach is based around an interoperability service which keeps details of all autonomic managers present and maintains a mapping of the resources they manage and their scope of operation and management. Autonomic managers register with the service via a standard interface (much like a name service) and provide details of their management capabilities using a standardised description language. The interoperability service contains the logic to detect conflicts and when necessary send a signal to one of the involved managers to stop its management activity. This approach can be highly scalable and robust if the service is itself distributed and operates hierarchically with a dynamically elected global instance.

We have adopted the second approach because it is scalable, generalisable, has low component-interaction complexity and has the advantage of not requiring further negotiation once a conflict has been detected.

IV. INTEROPERABILITY SERVICE

This section presents the architecture of an IS to facilitate exploration of the requirements identified above, and thus investigate the feasibility of a universal IS. By 'universal' it is meant that the architecture promotes a CORBA-like view of autonomic systems development, in which it is intended that any two autonomic managers that comply with the architecture specification will be guaranteed to co-exist in a system, without undesirable interactions leading to instability.

The IS maintains a database of all registered AMs along with a mapping of the resources they manage and their scope of operation and management. AMs register with the service via a standard interface and provide details of their management capabilities using a standardised description language. The IS detects potential conflicts and sends appropriate signals to one or more AMs to e.g., stop, suspend or restrict their management activity. The strengths of this approach are that it is scalable, generalisable, has low component-interaction complexity and because conflict management is handled within the IS, the AMs are not involved in negotiation with peers.

The service has a hierarchical structure for scalability, enabling conflict detection at both global level (such as system-wide security management) and local level (such as platform-wide, or VM-wide, resource management) with respect to a particular AM. Additional levels can be added, with a communication infrastructure resembling that of a typical hierarchical service such as DNS.

It is important that conflict-detection is performed at the correct level. For example, an autonomic VM scheduler only has a potential conflict with an autonomic memory manager, if they are both operating on the same processor unit.

Figure 1 shows the system-level view. The IS comprises a number of service instances distributed throughout a

system. Each instance of the IS provides service to a local group of AMs, resolving conflicts that occur at the local level. One of these instances is dynamically elected to serve as the global instance, and deals with resource conflicts at system level.

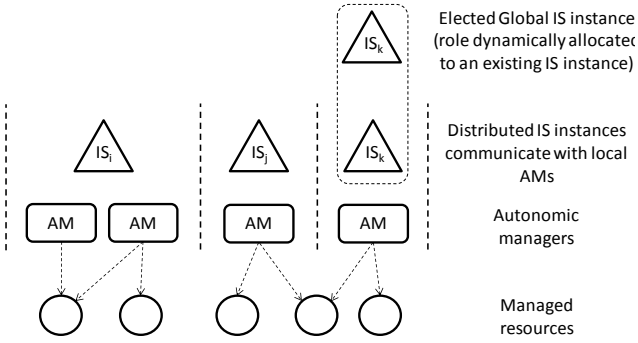


Figure 1. System-level view

The architecture is formed around a number of regular interfaces and a communication protocol which define the interaction between the components of the system, as shown in Figure 2.

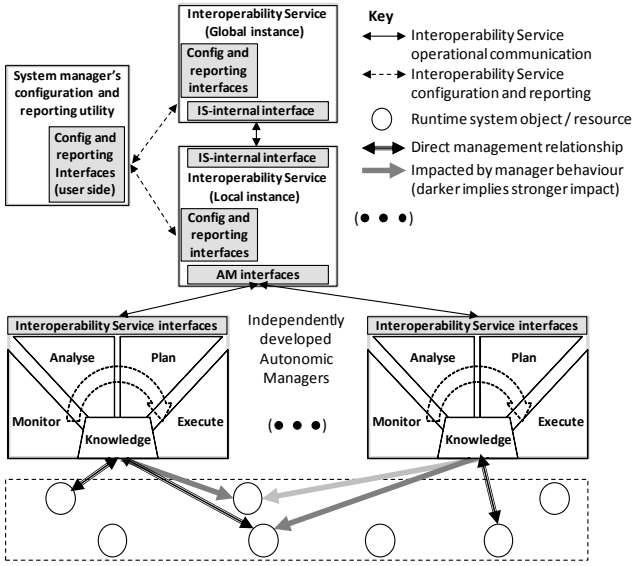


Figure 2. The Interoperability Service (IS) architecture, showing interface details.

A. Interoperability Service Interfaces

A number of interfaces are specified, and form three groups:

1. IS-AM interaction is supported by two interfaces.
   IAdvertise {*Advertise*, *Unregister*, *Heartbeat*} is used by AMs to signal joining (registering), leaving and heartbeat messages to the IS. *Advertise* is accompanied by a list of resources that the AM either wishes to manage directly, or that the developer has identified might be impacted by the manager's behaviour. This has the effect of registering the management interests of the AM with the IS. *Unregister* is used by an AM to signal an orderly shutdown, and *Heartbeat* (invoked periodically under normal conditions) enables (when absent) the IS to detect when a manager crashes or leaves abruptly. In either case, the AM's management interests are unregistered from the IS and the conflict detection analysis is triggered, so that any AMs which were suspended but are no longer in conflict with the system can be resumed.
   IInteroperate {*Run*, *Stop*, *Suspend*, *Restrict*, *Resume*, *Throttle*} is used to receive directives sent from the IS. The AM developer uses the IS API to map these directives onto the AM-internal behaviour. *Run* is accompanied by a sub-list of the requested resources that the AM can manage, so partial conflicts can be handled without suspending the entire manager. *Stop* shuts down the AM. *Suspend* backgrounds the AM (the AM developer determines the actual AM-internal semantics). *Restrict* is used to partially suspend an AM where potential conflict is discovered for a subset but not all of its management activities and is only used when the IS is configured to operate in the SAFE_COEXISTENCE mode (see later). *Resume* reactivates a suspended AM. *Throttle* provides for a more-sophisticated adjustment of AM behaviour in which the IS can specify different rates of management activity to potentially conflicting AMs to prevent certain oscillatory patterns developing.
2. IS-IS interaction is facilitated by a single interface.
   ICommunicate {*Forward*, *Locate*, *Elect*, *SetISLevel*, *GetISLevel*} supports hierarchical operation, necessary in large or complex systems when AMs operate at different levels within a system and may be involved in local or system-wide conflicts. *Forward* is used to pass messages between the Global IS instance and local ISs which want to control or impact on global-level resources (e.g., communication between low and high level scheduling managers); this is the basis of system-wide and cross-level conflict detection. The remaining functions support the hierarchical IS structure itself including leader election for robustness. *Locate* returns the ID of the current service coordinator IS instance (which also performs the role of global conflict detection). *Elect* initiates an election if no coordinator instance is found. *SetISLevel* is used to set the IS level status to be either Local or Coordinator. *GetISLevel* is used by each IS instance to determine its status during *Locate* and *Elect* events.
3. The IS provides an external management interface.
   IConfigure {*SetMode*, *GetMode*, *SetSensitivity*, *GetSensitivity*, *StatusReport*} is a configuration and reporting interface which allows external system management utilities to perform system-specific configuration and generate status reports and statistics. *SetMode* and *GetMode* allow run-time configuration of the service to allow different levels of safety; 'SAFETY_CRITICAL' requires that all of a particular AM's management activity is suspended when it is

found to be involved in a conflict, whilst 'SAFE_COEXISTENCE' allows partial suspension of AM functionality, such that only non-conflicting management activities continue. The IS is initialised to SAFETY_CRITICAL mode. *SetSensitivity* and *GetSensitivity* are used to configure the conflict detection sensitivity level (see section IV, part D) and to dynamically adjust this if necessary. *StatusReport* collects status information and statistics for report generation and IS performance monitoring.

The IS architecture specification defines the interfaces, and with its accompanying communication protocol, defines the message formats and sequences that form the inter-component communication. It also specifies the semantics of this communication. Figure 3 shows how the IS functionality is integrated with the various components of the system.

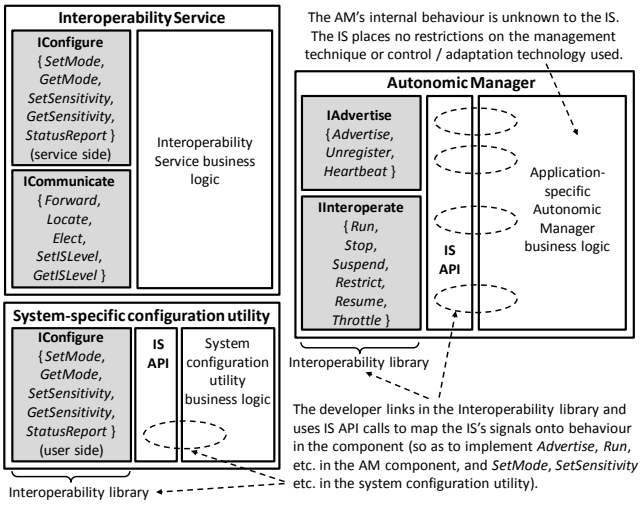


Figure 3. Internal architecture of the system components and the integration of the IS interfaces with these components.

The software developer retains flexibility with respect to the internal design and behaviour of the business logic of AM components and system configuration utilities. The architecture specification does not restrict the management approach, internal structure or control / adaptation techniques used within an AM component. However, the AM developer must integrate the API calls into the manager such that the control behaviour meets the IS specification (i.e., to interpret the directives {stop, suspend etc.} so that the AM's behaviour adheres to the respective IS semantics). Where an AM manages multiple resources the developer can choose to implement *Restrict* such that it is effective at the level of the AM itself, or only on the management activity that has been notified as being in conflict. In contrast *Suspend* always acts at the level of the entire AM. Similarly, the developer can decide the AM-internal semantics of *Suspend* and *Restrict* so as to isolate the management output (effecter output) of the manager whilst still running the monitor, analyse and plan parts if desired. This approach facilitates the IS' regulatory control over the AM when conflicts occur, whilst enabling 'warm' start-ups of components when conflicts are resolved.

B. The IS AM-state model

The IS maintains an instance of a state model for each locally registered AM (see Figure 4). The information held in these models drives the IS conflict management behaviour and is the basis on which AMs' management rights are governed.

An AM is discovered when it registers its management interests with its local IS instance. If there are no other AMs registered the new AM is granted management rights for the resources requested and signalled that it can run. If other AMs are already registered, the IS evaluates whether or not there is a possible conflict of interest, and if so signals the AM to either Stop (in which case the AM must attempt re-registration at a later time driven by some external event) or Suspend (in which case the IS will automatically signal the AM that it can resume, i.e., manage, once the conflict has been resolved).

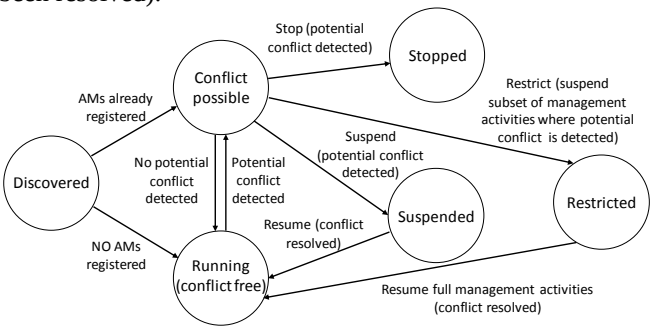


Figure 4. State diagram held by an IS instance, for each locally registered AM.

C. A Management Description Language

We discuss the need for a standard description of AMs' management interests, and briefly introduce our current language which is extensible to accommodate improvements in our understanding of ways actual and potential conflicts arise.

The IS facilitates interoperability (in the most limited case: safe coexistence) amongst (unknown in advance) AMs which have been developed independently of each other, and thus do not directly support interoperability amongst themselves.

The overall goal is to maximise the management freedom of AMs whilst at the same time ensuring that the system remains stable. To fulfil its main role, the IS must also:

- Detect AMs and learn their characteristics (via AM registration);
- Identify situations where conflicts can potentially occur, determine the consequences and the level of risk, and

achieve a system-specific balance when taking decisions to resolve conflicts by restricting, suspending or stopping AMs' management activities;
- Automatically enable the not-in-conflict subset of management activities for *restricted* AMs;
- Automatically resume suspended AMs when conflicts are resolved (e.g., on the basis or re-evaluating potential conflict status when other AMs leave the system);
- Enable cooperation between AMs. For example to share learnt knowledge concerning system state, volatility etc.

To perform these functions, the IS needs certain information detailing each AMs' management domain and specific resources of interest. This information must use a standard language format, and a fixed vocabulary of key terms so that automated searching for overlaps of interest can be performed effectively. The information will be provided at run time by the AM via the IS API (the information is provided ultimately by the AM developer).

Conflicts can arise in several ways. Direct conflicts occur where multiple AMs attempt to manage the same resource or object. However conflicts can be indirect (and less obvious) because a manager's activity may impact resources other than those directly managed. Categories of this include *cross-application* conflicts, for example increasing a specific application's use of a particular resource such as network bandwidth reduces the availability of bandwidth available to other applications. Another category of indirect conflicts are *cross-resource* conflicts, for example increasing processor speed to maximise throughput increases direct power usage and may also increase power requirements for cooling systems (which may have their own autonomic management systems). Some system characteristics such as security policy, power usage, server provisioning strategy etc. may be managed at both the system-wide level, and locally at the level of individual computing node or cluster. This can lead to conflicts between global and local managers, resulting in parts of the system being out-of step with global policy, and/or inefficient behaviour.

Clearly, it is difficult to identify every possible case of indirect conflict with certainty, and the *extent* of management impact in such cases is also highly variable. Therefore the description information provided by AMs must be sufficient to derive a similarity measure between their management effects. The language needs to contain appropriate categories to express areas of management concern in a structured way, i.e., from high-level domain in which the manager operates down to specific resources that are managed, and also to express characteristics including the management scope (global or local) and specificity (e.g., organisation specific, application specific).

Given these requirements, the standard management description should include:

Category. Mandatory. The highest-level and most generic descriptor used to identify the AM's domain of interest. Terms include:
{*Power general*, *Performance general*, *Security general*, ...}

Zone. Mandatory. A second level, more specific sub-category enabling developers to differentiate between specific management functions. Terms include:
{*Power system*, *Power platform*, *Power cooling* ... *Performance system*, *Performance CPU*, *Performance disk*, *Scheduling*, *VM management*, ... }

Impact. Mandatory. A numerical indicator Impact Factor (IF), (where $0 < IF \leq 1$), is defined to express the strength of the management influence. A directly controlled resource or parameter is assigned the value 1. A value close to 0 indicates that the particular AM has a weak influence on the resource whilst values close to 1 indicate that the resource is closely impacted by changes to one that is directly managed by the AM. For example an AM directly controlling CPU speed (IF = 1) has a strong indirect influence on VM performance ($IF \approx 0.8$). Term: { *ImpactFactor*(value) }

Scope. Mandatory. Whether the manager has local or global impact. Terms: { *Local*, *Global* }

Specificity. Optional. The extent of manager operation. Terms include: { *System-wide*, *Application-wide*, *Platform-wide*, *Process-wide*, *User-specific*, ... }

Trigger. Optional. This facilitates expression of temporal aspects such as periodicity or operating timescale, as well as specific events that invoke the management activity. Such characteristics can potentially be used to detect combinations of AMs at risk of causing of instability in the form of oscillation or control divergence for example. Terms include: { *Period*(value), *Event*(name) , ... }

Parameter. Optional. Identification of specific context parameters that are of interest to the AM. Term:
{ *Name*(value) }

Envelope. Optional. Expression of range of control freedom for a given named Parameter. This can potentially help to avoid false positive detections of conflict, when managers operate in the same domain but have non-overlapping envelopes of operation. Terms include:
{ *Name*(range, value) }

Where provided, the Envelope term allows more precise determination of the risk of conflict in cases where a pair of AMs both declare an envelope value for a specific parameter. Where an AM does not declare an envelope value for any given Parameter the full state space of values is assumed.

D. Conflict Detection

The architecture specification does not mandate the actual conflict detection technique to be used; this is an

implementation decision and will be based on the level of sophistication required in a particular system.

In our exploratory work conflict detection is based on calculating a numerical measure of similarity between the management interests of a pair of AMs, and comparing this measure with a sensitivity threshold level. A newly registering AM's management description is compared with those of the already registered AMs.

The technique is described below and an example implementation is outlined in section V.

The architecture specification defines a dynamically configurable conflict sensitivity threshold ($0 < Thresh_C \leq 1$) which is used to tune the conflict detection sensitivity (via *SetSensitivity*, on **IConfigure**). A potential conflict is detected if the similarity match measure *Match* of a pair of AMs exceeds $Thresh_C$. The sensitivity level is configured by the facility manager via a control console application (or tuning of this parameter could be automated), and can be changed at run time as necessary. This enables safety critical systems (for example) to operate pessimistically with very low tolerance to potential manager conflicts, whereas in domains where only efficiency (for example) is at stake, the system can operate more optimistically, with a higher tolerance which can lead to benefits of having a greater number of AMs working simultaneously (bearing in mind that a 'potential conflict' may not be realised).

V. IMPLEMENTATION

This section describes a work-in-progress implementation which employs a subset of the extensible architecture's characteristics for demonstration of the core behaviour. Here we focus on the operation of the service at a local level, since it is intuitive to expect that many conflicts between autonomic managers will be localised due to decisions concerning local resources, or configurations of local services.

The IS maintains a table which contains the identity and state of each registered AM, and a second table which keeps track of each AM's directly managed and indirectly impacted resources (see figure 5). Information in this table comprises: AM_ID (a value allocated to the AM by the IS during the discovery process); General area of management function (a 'category' term from the management description language); Sub-classifier of management function (a 'zone' term from the management description language); Managed parameter name ACItem_ID (the optional 'parameter' term from the management description language); Conflict status and Impact Factor for the related resource; and Scope (a 'scope' term from the management description language). Figure 5 also shows the communication that takes place between an AM and the IS. MAdvertise, MRelease and MHeartbeat are messages sent from the AM via actions on the IAdvertise interface. MACK / MNACK are Acknowledge / Not Acknowledge responses to management requests accompanying MAdvertise. This works as follows: the AM tries to register (Advertise) its management interests one by one and the IS replies with MNACK messages if any are in conflict with the rest of the system, MACK otherwise. MSuspend, MResume, MRun, MStop and MThrottle are directives sent by the IS via the IInteroperate interface.

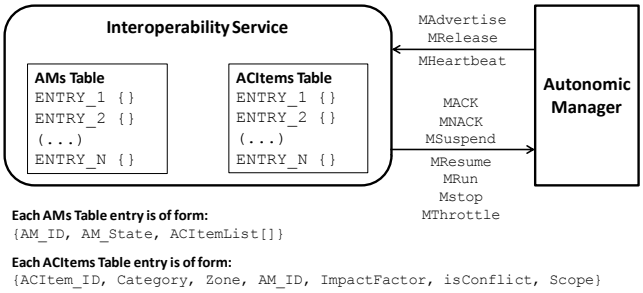


Figure 5. The IS' internal data tables, and overview of the AM-IS communication protocol.

For initial exploration we use a conflict detection technique based on a numerical similarity measure of AMs' management interests. Conflict detection activity is triggered by events that change the population or configuration of the AMs; such as the registration of a newly-discovered AM, or the departure of an AM from the system.

For a pair of AMs $\{AM_i, AM_j\}$ the similarity measure $Match_{ij}$ is derived from the management descriptions of the AMs as follows:

Let N_i = name of the specific managed resource (specified by the Parameter term in the management description),

C_i = management category,

Z_i = management zone,

IF_i = impact factor (of AM_i on the resource identified by $\{N_i, C_i, Z_i\}$),

S_N, S_C, S_Z = similarity indicator of management description terms Name, Category and Zone respectively for the pair of AMs.

$$Match_{ij} = \frac{S_N + S_C + S_Z + IF}{4}$$

where:

$$S_N = \begin{cases} 1 & \text{when } N_i = N_j \\ 0 & \text{when } N_i \neq N_j \end{cases},$$

$$S_C = \begin{cases} 1 & \text{when } C_i = C_j \\ 0 & \text{when } C_i \neq C_j \end{cases},$$

$$S_Z = \begin{cases} 1 & \text{when } Z_i = Z_j \\ 0 & \text{when } Z_i \neq Z_j \end{cases},$$

$$IF = \frac{IF_i + IF_j}{2}.$$

IF values are normalised, i.e., $IF_i, IF_j \in (0,1]$, thus the resulting similarity measure will always be a normalised value $Match_{ij} \in (0,1]$.

A newly registering AM's management interests are compared with details of each already registered AM, at the local IS instance in most cases. This is performed independently for each resource pair combination; so if AM_i and AM_j are registered with declared management interests in *m* and *n* resources respectively, and AM_k attempts to register *p* resource management interests, then *mp* + *np* similarity measures are generated.

A potential conflict is detected if for any pair of AMs {i,j}, $Match_{ij}$ exceeds the conflict sensitivity threshold ($Thresh_C$).

When evaluating the scalability of the approach it is important to consider: 1. conflict detection occurs predominantly at the level of the local IS instance; only in cases where an AM's resource description has global scope does the conflict detection get invoked at the global level; 2. conflict detection is only performed when events that affect the AM population occur (e.g., AMs arriving, leaving); and 3. whilst we do not limit the number of AMs registered at a local IS instance, we expect this number to be of order 10, or perhaps 100 rather than much bigger values, for realistic systems.

The dynamically configurable operating mode of the IS determines what action is taken once a potential conflict has been detected. If the IS mode is SAFETY_CRITICAL, AM_k will be *suspended* (i.e., management activities are inhibited at the level of the AM itself). In SAFE_COEXISTENCE mode AM_k will be *restricted*, (i.e., management activities are inhibited at the level of specific resources managed by a particular AM; it is allowed to perform its normal management operations for the not-in-conflict subset of its management domain). The actual semantics for restricted AM-internal operations are to some extent implementation specific. In some cases it will be desirable to enable the *monitoring* aspect to operate as normal (to prevent discontinuity in monitoring traces etc., and to facilitate warm restarts of restricted operations), but in all cases the *effecter* is switched off, i.e., the manager can monitor its environment but cannot change anything.

The current implementation uses policy-based management logic within AMs; and is based on Agile++ [16], [17]. Agile++ has language components including Rules, Variables and Actions. Under typical normal behaviour, a Rule will be evaluated to determine which Action needs to be performed, using Environment Variables to reflect external inputs to the Rule and Output Variables to signal the result of an Action. *Restricted* mode has been implemented for conflicting operations such that the AM still evaluates its control policy and executes Actions within, as normal. However, Output Variables are disabled (value forced to NULL) so that the Action can continue to make internal updates (such as for external-state tracking) but cannot actually effect the external system state.

As an alternative to using the IAdvertise interface for AMs to register their management interests, the implementation supports the encoding of the Management Description Language in XML format. An example configuration file is shown in Figure 6.

```
<!-- Autonomic Manager Configuration Specification Language -->
<MetaData>
  <ConfigAuthor Name="Mariusz Pelc" Organisation="UoG" />
  <TimeStamp Time="12:00" Date="20/12/2010" />
    <AMDescription>
      <AM ID="AM1">
        <ACItems>
          <ACItem ID="Performance" Scope="Local">
            <Category>Performance General</Category>
            <Zone>CPU Performance</Zone>
            <ImpactFactor>1.0</ImpactFactor>
          </ACItem>
          <ACItem ID="Power" Scope="Local">
            <Category>Performance General</Category>
            <Zone>System Performance</Zone>
            <ImpactFactor>0.5</ImpactFactor>
          </ACItem>
        </ACItems>
      </AM>
    </AMDescription>
</MetaData>
```

Figure 6. XML representation of the Management Description Language

A. Wider Architectural Perspective

The IS implementation forms part of a wider project to develop a full component model and middleware for autonomic computing which has been ongoing at Greenwich for several years, see for example [18], [19]. Full details of this are out of scope for this paper, but in brief, this is a policy-based system in which services including communication manager, context manager, repository manager and now the IS are optionally policy supervised. The middleware supports policy-based application-specific components which can have dynamic (run-time) policy upgrades and which have in-built fault recovery. For example if a new policy is loaded but its required context information is not available from the context manager then an automatic roll-back to a previously working policy is performed. Architectural support for low-resourced embedded platforms is also included.

B. Evaluation Application Scenario

Data centre management is a popular application domain for AC; due in part to the high configuration complexity that arises from the scale of operation, and also because with such large amounts of resources deployed the potential efficiency savings are very high. AC currently targets several key aspects of data centres, including power management to reduce running costs, and scheduling to improve resource efficiency. We demonstrate the operation and benefit of the IS in a data centre scenario in which two independently developed AMs coexist (managing power usage, and processor scheduling, respectively); their management operations potentially conflicting.

The scenario: The scheduling manager (AM1) has a main goal of maximising throughput by keeping all resources utilised where possible. The power manager (AM2) is designed to minimise power usage by slowing down processor speed or by shutting down entire processor units where possible. We assume that, in the absence of other managers, each of these services has been extensively evaluated and found to improve overall performance.

The co-existence of these AMs creates a high potential for conflict. For example AM2 will attempt to shutdown an underutilised resource as soon as load level starts to fall, whilst AM1 will attempt to bring unused resources into play as soon as load levels increase (or a backlog develops). Depending on the sequence of load level changes it is possible that oscillation will build up between the actions of these two managers.

Operation: During its initialisation each AM registers with the IS. The management capabilities of each AM are described using the standard language and categories described earlier.

AM1 directly controls a parameter *performance* within the general management category *performance general*, and specific sub-zone *CPU performance*; and indirectly influences a parameter *power* within the general category *performance general*, and sub-zone *system performance*.

AM2 directly controls a parameter *power* within the general category *power general*, and the specific zone of interest *system power*; and indirectly influences a parameter *performance* within the general category *performance general*, and the specific zone of interest *CPU performance*.

```
a) AddACItem     ("Performance", "Performance General",
                  "CPU Performance", "1.0", "Local");
   AddACItem     ("Power", "Performance General",
                  "System Performance", "0.5", "Local");
   RegisterAsAM  ();

b) AddACItem     ("Power", "Power General",
                  "System Power","1.0","Local");
   AddACItem     ("Performance", "Performance General",
                  "System Performance", "0.5", "Local");
   RegisterAsAM  ();

c) bool AddACItem(char *ParameterName, char *Category,
                  char *Zone, char *Impactfactor, char *Scope);
```

Figure 7. API calls to register AM's management interests.

The API calls to perform the manager registration with the IS are shown in Figure 7a (for AM1), and 7b (for AM2), where AddACItem means 'Add autonomically controlled item'; its template is shown in Figure 7c.

VI. EVALUATION

As mentioned in section V, part A this work forms part of a larger project to develop a full component model and middleware for autonomic computing. We use the existing infrastructure as a testbed to evaluate the IS in a realistic system setting.

In addition to the IS, three additional system services are provided to create a run-time environment in which the behaviour of the IS and AMs can be evaluated, these are: Communication Manager; ContextManager and RepositoryManager. In addition, a couple of services were fabricated to provide mock context values for two system parameters which are needed as inputs in the run-time execution of various control policies used in the experiments. The EfficiencyProvider component generates the 'Efficiency' parameter, and likewise the LoadProvider component generates the 'Load' system parameter.

The services are integrated into a middleware component (available in the form of shared library for Linux) with API interface enabling communication, context and repository management, conflict resolving and policy evaluation.

Two IS-compliant AMs (AM1, AM2) have been developed to evaluate and demonstrate the behaviour of the Interoperability Service. AM1 and AM2 target popular management domains within cloud / grid computing, typical of autonomic control systems currently deployed in data centre systems for example. The whole application (including the AMs) thus comprises of 8 services. Figure 8 provides a snapshot of the system in operation during scenario 5 (see below), showing clockwise from top left: Communication Manager, Context Manager, Interoperability Service, AM2, AM1, and the Repository Manager.

The management domains of AM1, AM2 respectively are: processor scheduling (with the goal of maximising throughput by keeping resources utilised where possible), and power management (with the goal of minimising power usage). This is a realistic situation in which the direct management activities are well differentiated, but in which there is an indirect conflict as discussed in section IV, part C.

The AMs are designed so as to be representative of *independently developed* components operating in a data-centre system, i.e., the AMs include no direct support for co-existence or interoperability amongst themselves. The evaluation is performed in 5 scenarios. The first four scenarios show the behaviour of the IS when operating in SAFETY-CRITICAL mode under a range of different resource management circumstances. The fifth scenario shows how the IS responds to AM conflicts when the IS is operating in SAFE-COEXISTENCE mode.

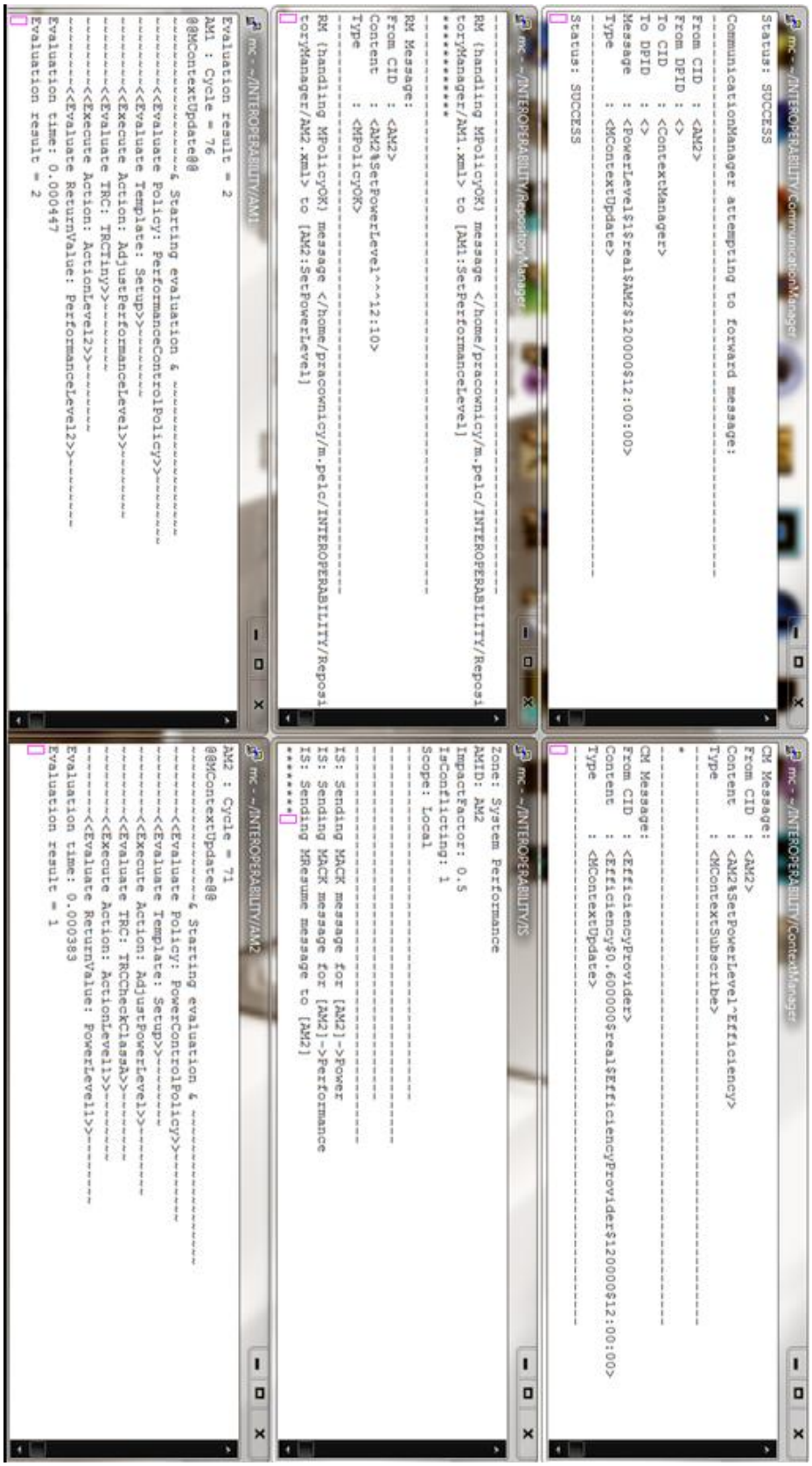


Figure 8. The system in operation during the evaluation.

Scenario 1 illustrates the standalone manager case, and is included for completeness. Each manager registers separately in the system in the absence of the other. $Thresh_C$ = 0.6. AM1 requests management rights for CPU performance, and also notifies a potential impact on system power. As there are no other AMs present, the IS grants AM1 permission to manage unimpeded. Similarly, for AM2 (in the absence of AM1) the IS grants rights to manage system power level and also to have an indirect impact on system performance.

Scenario 2 illustrates the case where a potential conflict is detected between a pair of managers (IS operating in SAFETY-CRITICAL mode). AM1 registers with the IS and is granted rights to manage the resources it has requested. AM2 then registers whilst AM1 is still present. $Thresh_C$ = 0.6. The IS performs conflict detection analysis, based on the AMs' announced Impact Factors (IFs) for each requested managed item. This determines whether AM2 can be granted the requested management rights: *Power* directly managed (IF=1.0), and *Performance* potentially affected indirectly (IF=0.5). The match levels are determined using the algorithm presented in section V. In this case a conflict is detected; arising from AM1's direct management of performance and AM2's indirect impact on performance, giving a match value greater than the threshold. This can be seen in the diagnostic trace in figure 9.

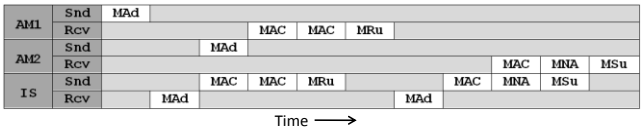
```
IS: Handling Advertise Message:
IS: Conflict Detection [AM2->Power]::[AM1->Performance]
IS: Match Level=0.25, Threshold=0.6
IS Decision: No Conflict Detected
IS: Conflict Detection [AM2->Power]::[AM1->Power]
IS: Match Level=0.4375, Threshold=0.6
IS Decision: No Conflict Detected

IS: Conflict Detection [AM2->Performance]::[AM1->Performance]
IS: Match Level=0.6875, Threshold=0.6
IS Decision: Conflict Detected
IS: Conflict Detection [AM2->Performance]::[AM1->Power]
IS: Match Level=0.625, Threshold=0.6
IS Decision: Conflict Detected

IS: Sending MACK message for [AM2]->Power
IS: Sending MNACK message for [AM2]->Performance
IS: Sending MSuspend message to [AM2]
```

Figure 9. A potential conflict is detected.

Figure 9 shows a diagnostic trace of the IS conflict detection process, in which the advertised management interests of AM2 are compared for all relevant AMs. In this specific case AM1 is already managing a system performance characteristic (specifically CPU performance), when AM2 registers, requesting to manage system power, but also announcing a potential impact on system performance. The IS does not detect a direct conflict with the power management, but the conflict match level for system performance exceeds the current $Thresh_C$ (0.6). The IS suspends the newly registering manager to prevent possible instability (this manager will be automatically resumed if AM1 leaves the system and there are no other conflicts with other AMs registered in the meantime). Figure 10 shows the resulting message sequence.




```
Key:  Snd - Sent Message          MNA - MNACK          MRu - MRun
      Rcv - Received Message      MRl - MRelease       MSp - MStop
      MAd - MAdvertise Message    MRe - MResume
      MAC - MACK Message          MSu - MSuspend
```

Figure 10. Message sequence for scenario 2.

Scenario 3: As scenario 2, but with $Thresh_C$ = 0.8, i.e., the IS is less sensitive to potential conflicts (this configuration may be better suited to non-critical systems where some potential for conflict may be acceptable, i.e., the tradeoff between safety and management flexibility is shifted). The new diagnostic behaviour trace and the resulting message sequence are shown in Figure 11 and Figure 12 respectively. In this case no conflicts are detected and the newly arriving AM2 is granted rights to manage system power level, and to have an impact on system performance, thus potentially interacting with AM1.

```
IS: Handling Advertise Message:
IS: Conflict Detection [AM2->Power]::[AM1->Performance]
IS: Match Level=0.25, Threshold=0.8
IS Decision: No Conflict Detected
IS: Conflict Detection [AM2->Power]::[AM1->Power]
IS: Match Level=0.4375, Threshold=0.8
IS Decision: No Conflict Detected

IS: Conflict Detection [AM2->Performance]::[AM1->Performance]
IS: Match Level=0.6875, Threshold=0.8
IS Decision: No Conflict Detected
IS: Conflict Detection [AM2->Performance]::[AM1->Power]
IS: Match Level=0.625, Threshold=0.8
IS Decision: No Conflict Detected

IS: Sending MACK message for [AM2]->Power
IS: Sending MACK message for [AM2]->Performance
IS: Sending MRun message to [AM2]
```

Figure 11. IS conflict detection analysis in which the conflict match level is below the conflict threshold.

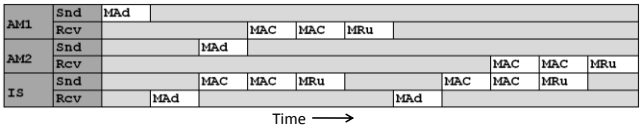


Figure 12. Message sequence for scenario 3.

Scenario 4 Illustrates the case where AMs are replicated and the IS must ensure that only a single instance is active at any time (note that the IS does not know that the two managers are identical, it bases its decisions only on the AMs' management descriptions). Manager AM1 registers and begins managing its advertised resource. A second instance of the *same manager type as AM1*, AM3, requests management rights from the IS. $Thresh_C$ = 0.6. The conflict detection procedure is not executed when AM1 registers as there are no other AMs registered with the IS. Thus AM1 is granted management rights for both resources requested. The registration of AM3, advertising a direct management interest in *Performance* and an indirect impact on *Power*, triggers conflict detection analysis, as shown in Figure 13.

In this case, conflicts are detected for both of the requested resources, so as a result, AM3 is suspended. At a later time, AM1 performs an orderly shutdown sending an

MRelease message to the IS, invoking the UnregisterAM function at the IS. This has 3 effects: 1. an *MStop* message is sent to AM1 (see Figure 14); 2. the IS unregisters all AM1's management interests; 3. conflict detection analysis is again triggered, now with the goal of detecting situations where previous conflicts have now been resolved. Any suspended AM's that are no longer in conflict with active managers are now resumed. In this case AM3 is the only suspended AM, and in the absence of any conflicts with active AMs it is automatically resumed and granted its requested management rights (see Figure 15).

```
IS: Handling Advertise Message:
IS: Conflict Detection [AM3->Performance]::[AM1->Performance]
IS: Match Level=1, Threshold=0.6
IS Decision: Conflict Detected
IS: Conflict Detection [AM3->Performance]::[AM1->Power]
IS: Match Level=0.4375, Threshold=0.6
IS Decision: No Conflict Detected

IS: Conflict Detection [AM3->Power]::[AM1->Performance]
IS: Match Level=0.4375, Threshold=0.6
IS Decision: No Conflict Detected
IS: Conflict Detection [AM3->Power]::[AM1->Power]
IS: Match Level=0.875, Threshold=0.6
IS Decision: Conflict Detected

IS: Sending MNACK message for [AM3]->Performance
IS: Sending MNACK message for [AM3]->Power
IS: Sending MSuspend message to [AM3]
```

Figure 13. Conflict detection analysis finds potential conflicts of interest between two instances of the same AM type.

```
IS: Handling Release Message:
IS: Sending MStop message to [AM1]
```

Figure 14. IS receives *MRelease*, responds with *MStop*.

```
List of Suspended AMs:
-----------------------------------------------
AM Name: AM3
AM State: SUSPENDED
-----------------------------------------------
IS: Sending MACK message for [AM3]->Performance
IS: Sending MACK message for [AM3]->Power
IS: Sending MResume message to [AM3]
```

Figure 15. IS resumes the AM3 Manager

Figure 15 illustrates the IS's behaviour on receipt of an *MRelease* message, which implies that an AM has left the system and thus one or more previously detected conflict conditions may have been removed. First the state model is searched for any AMs in the SUSPENDED state. The management interests of these are re-examined against those of the remaining RUNNING state AMs (conflict detection analysis is triggered again). Any suspended AMs which are now conflict-free are resumed (AM3 in this case). Figure 16 shows the entire message sequence for scenario 4.

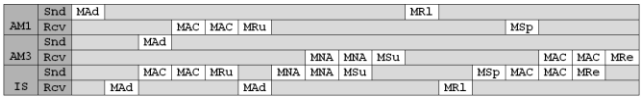


Figure 16. Message sequence for scenario 4.

In addition to illustrating the prevention of conflicts of directly overlapping management interest; scenario 4 also shows how the IS architectural approach facilitates and manages redundant replication of autonomic manager processes for robustness within a system. Only one AM is given management rights for a particular resource at any time, but whenever an AM leaves the system the set of running and suspended AMs is automatically re-evaluated for changes in conflict status. Suspended replicas are resumed when determined conflict-free, and can start 'warm' because the AM's developer can choose to implement '*suspend*' as only shutting down the execute stage of the MAPE loop.

Scenario 5 is the equivalent of scenario 2, except that in this case the IS operates in SAFE-COEXISTENCE mode. AM1 registers its management interests with the IS, followed by AM2. $Thresh_C$ = 0.6. The two Autonomic Managers attempt to control respectively, Performance (direct control with IF=1.0) and Power (indirect control with IF=0.5) for AM1 and Power (direct control, IF=1.0) and Performance (indirect, IF=0.5) for AM2.

As there are no other AMs running when AM1 registers it is granted full management rights, as shown in figure 17.

```
IS: Handling Advertise Message:
IS: Sending MACK message for [AM1]->Performance
IS: Sending MACK message for [AM1]->Power
IS: Sending MRun message to [AM1]
```

Figure 17. IS issues full rights to the AM1 Manager

When AM2 registers its management interest the IS checks for a conflict with all other registered managers. As a result the IS allows AM2 to control Power but **restricts** controlling Performance and sends an MRestrict message to AM2 as the diagnostic trace in figure 18 shows.

```
IS: Handling Advertise Message:
IS: Conflict Detection [AM2->Power]::[AM1->Performance]
IS: Match Level=0.25, Threshold=0.6
IS Decision: No Conflict Detected
IS: Conflict Detection [AM2->Power]::[AM1->Power]
IS: Match Level=0.3875, Threshold=0.6
IS Decision: No Conflict Detected
IS: Conflict Detection [AM2->Performance]::[AM1->Performance]
IS: Match Level=0.6875, Threshold=0.6
IS Decision: Conflict Detected
IS: Conflict Detection [AM2->Performance]::[AM1->Power]
IS: Match Level=0.575, Threshold=0.6
IS Decision: No Conflict Detected
IS: Sending MACK message for [AM2]->Power
IS: Sending MNACK message for [AM2]->Performance
IS: Sending MRestrict message to [AM2]
```

Figure 18. A potential conflict is detected; AM2 is restricted.

In the Restricted mode AM2 evaluates its policy as normal but the Performance output variable is set to NULL, i.e., AM2 cannot actually effect the system performance whilst restricted in this management aspect. AM2 manages power normally, as this aspect was not restricted.

Later, AM1 Unregisters with the IS, this again triggers conflict check operation. AM2 is no longer in conflict, so is now granted permission to control all items of interest, as shown in the trace in figure 19.

```
IS: Handling Release Message:
delete AMDesc: AM1
IS: Sending MStop message to [AM1]
List of Restricted AMs:
-----------------------------------------------
AM Name: AM2
-----------------------------------------------
ACItem Name: Power
Category: Power General
Zone: System Power
AMID: AM2
```

```
ImpactFactor: 1.0
IsConflicting: 0
Scope: Local
----------------------------------------
----------------------------------------
ACItem Name: Performance
Category: Performance General
Zone: System Performance
AMID: AM2
ImpactFactor: 0.5
IsConflicting: 1
Scope: Local
----------------------------------------
------------------------------------------------
IS: Sending MACK message for [AM2]->Power
IS: Sending MACK message for [AM2]->Performance
IS: Sending MResume message to [AM2]
```

Figure 19. Diagnostic trace showing IS behaviour during release of AM1 and subsequent granting of full rights to AM2

Figure 20 shows the entire message sequence for scenario 5.

A. Evaluation summary

The evaluation aspect of this paper is mainly concerned with demonstration of the IS concept. Our implementation is not necessarily optimised for processing performance. We have focussed the evaluation on exploring the behaviour of the system in a set of base cases which represent realistic types of conflicts (direct and indirect) between AMs.

The evaluation was based on a number of ways in which a pair of AMs may overlap in their management activities. These scenarios were chosen so as to reflect a wide range of possibilities.

The case results show how the IS controls the management rights of AMs dynamically, using the management similarity-measure based conflict detection. We have demonstrated the variable safety-sensitivity of the service, using the configurable sensitivity threshold combined with the choice of two safety levels (SAFTEY-CRITICAL and SAFE-COEXISTENCE).

The processing overhead of conflict detection does not increase significantly when larger populations of AMs exist, because conflict detection is only triggered when the AM population changes (e.g., a new AM is registered), and the existing AMs are only compared against the arriving AM (not against each other). The conflict detection always considers AMs on a pair-wise basis, so functional complexity remains the same regardless of the number of AMs present.

VII. CONCLUSION

In this paper, we have outlined the case for greater research effort in the area of interoperability of autonomic managers. We have discussed why bespoke and custom solutions will not work in the long term and argued for a universal standard for interoperability. In line with this we have identified requirements for a service-based approach.

We are working towards standards and services for universal interoperability in autonomic systems. In particular we are targeting the under-addressed challenge of interoperability and co-existence in not-planned circumstances, i.e., for AMs that are developed independently and brought together when systems are built from a number of separate components, and also when existing systems or components are upgraded.

This work is timely and important because the likelihood of conflicts will escalate as autonomic computing continues to increase in popularity, and AMs are deployed in an ever-wider array of components with ever-richer functionality.

We have presented initial work towards a service-based automatic and proactive interoperability service, being integrated into autonomic components and making them ‘interoperability ready’ in advance of their deployment. Our approach enables AMs to be developed independently, requiring that the developer uses a management description language to describe the component’s management characteristics. This approach has the main advantage of not requiring an AM developer to have knowledge of other AM’s that may exist in the target system, and thus supports agility i.e., configuration changes, expansion and upgrades.

The technique has been developed with generalisation as a main goal. In the same way that it is not possible when developing an AM to perceive all the possible other AMs and their management capabilities that could coexist; it is also not possible when developing an IS to predict all of the application domains and behaviours of future AMs. Therefore we have ensured that the language used to describe management capabilities is extensible, and can be represented using a standard format (XML). The architecture defines the interfaces and communication between the key management components of the system but leaves open the implementation decisions for the IS-internal business logic so it can be tailored to a system’s needs.

The demonstration-of-concept implementation has focussed initially on ‘safe coexistence’ as a mandatory foundational step towards universal AM interoperability. Further work focuses on more-sophisticated techniques for the conflict detection, and further refinement of the management description language on which the conflict detection is based.

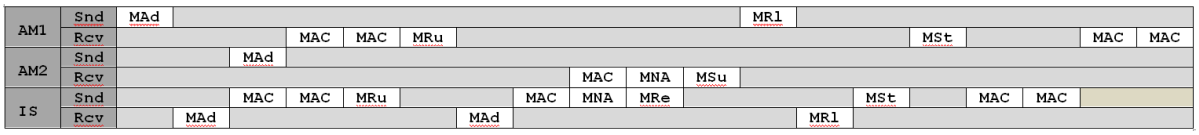


Figure 20. Message sequence for scenario 5.

VIII. REFERENCES

[1] Anthony R, Pelc M, and Shuaib H, The Interoperability Challenge for Autonomic Computing, The Third International Conference on Emerging Network Intelligence (EMERGING 2011), Lisbon, Portugal, November 20-25, 2011, pp. 13-19, IARIA, ISBN 978-1-61208-174-8.

[2] Kephart J. O., Chan H., Das R., Levine D. W., Tesauro G., Rawson F., and Lefurgy C. 2007. Coordinating multiple autonomic managers to achieve specified power-performance tradeoffs. In *Proc. 4th Intl. Conf. on Autonomic Computing* (Jacksonville, FL, USA, June 2007). ICAC'07. IEEE, 1–9.

[3] Wang M., Kandasamyt N., Guezl A., and Kam M. 2006. Adaptive performance control of computing systems via distributed cooperative control: Application to power management in computing clusters. In *Proc. 3rd Intl. Conf. on Autonomic Computing* (Dublin, Ireland, June 2006). ICAC'06. IEEE, 165–174.

[4] Zhao M., Xu J., and Figueiredo R. J. 2006. Towards autonomic grid data management with virtualized distributed file systems. In *Proc. 3rd Intl. Conf. on Autonomic Computing* (Dublin, Ireland, June 2006). ICAC'06. IEEE, 209–218.

[5] Khargharia B., Hariri S., and Yousif M. S. 2006. Autonomic power and performance management for computing systems. In *Proc. 3rd Intl. Conf. on Autonomic Computing* (Dublin, Ireland, June 2006). ICAC'06. IEEE, 145–154.

[6] Xu J., Zhao M., Fortes J., and Carpenter R. 2007. On the use of fuzzy modeling in virtualized data center management. Autonomic Computing, 2007. In *Proc. 4th Intl. Conf. on Autonomic Computing* (Jacksonville, FL, USA, June 2007). ICAC '07. IEEE Computer Society.

[7] Wang R., Kusic D. M., and Kandasamy N. 2010. A distributed control framework for performance management of virtualized computing environments. In *Proc. 7th Intl. Conf. on Autonomic Computing* (Washington DC, USA, June 2010). ICAC'10. IEEE, 89–98.

[8] Ghanbari S., Soundararajan G., Chen J., and Amza C. 2007. Adaptive learning of metric correlations for temperature-aware database provisioning. In *Proc. 4th Intl. Conf. on Autonomic Computing* (Jacksonville, FL, USA, June 2007). ICAC'07. IEEE.

[9] Kutare M., Eisenhauer G., and C. Wang. 2010. Monalytics: Online monitoring and analytics for managing large scale data centers. In *Proc. 7th Intl. Conf. on Autonomic Computing* (Washington DC, USA, June 2010). ICAC'10. IEEE, 141–150.

[10] Zhu X., Young D., Watson B. J., Wang Z., Rolia J., Singhal S., McKee B., Hyser C., Gmach D., Gardner R., Christian T., and Cherkasova L. 2008. 1000 islands: Integrated capacity and workload management for the next generation data center. In *Proc. 5th Intl. Conf. on Autonomic Computing* (Chicago, IL, USA, 2008). ICAC '08. IEEE, 172–181.

[11] Poussot-Vassal C., Tanelli M., and Lovera M. 2010. A Control-Theoretic Approach for the Combined Management of Quality-of-Service and Energy in Service Centres. In *Run-time Models for self-managing Systems and Applications*. Ardagna D and Zhang L, Eds). Springer Basel AG. 73-96.

[12] White S. R., Hanson J. E., Whalley I., Chess D. M., and Kephart J. O. 2004. An architectural approach to autonomic computing. In *Proc. 1st Intl. Conf. on Autonomic Computing* (New York, NY, USA, May 2004). ICAC'04. IEEE. 2-9.

[13] Kennedy C. 2010. Decentralised metacognition in context-aware autonomic systems: some key challenges. In *Proc. American Institute of Aeronautics and Astronautics* (AIAA) *Workshop on Metacognition for Robust Social Systems* (Atlanta, Georgia,) AAAI-10, AIAA. 34-41.

[14] Salehie M. and Tahvildari L. 2005. Autonomic computing: Emerging trends and open problems. In *Proc. Workshop on the Design and Evolution of Autonomic Application Software* (New York, NY, USA, 2005). DEAS'05. ACM Special Interest Group on Software Engineering. 30. 1–7.

[15] Quitadamo R. and Zambonelli F. 2008. Autonomic communication services: a new challenge for software agents. *SpringerLink Journal of Autonomous Agents and Multi-Agent Systems*. 17, 3 (2008), 457–475.

[16] Anthony, R. J. Policy-based autonomic computing with integral support for self-stabilisation, *International Journal of Autonomic Computing*, Vol. 1, No. 1, pp.1-33. ISSN (Online): 1741-8577, ISSN (Print): 1741-8569, 2009, Inderscience.

[17] P. Ward, M. Pelc, J. Hawthorne, and R. Anthony, Embedding Dynamic Behaviour into a Self-configuring Software System, *In Proc. 5th Intl Conf. on Autonomic and Trusted Computing* (ATC 2008), Oslo, Norway, Lecture Notes in Computer Science (LNCS 5060/2008), ISBN 978-3-540-69294-2, pp373-387, June 23-25, 2008, Springer-Verlag.

[18] Anthony R. J., Pelc M., Ward P., and Hawthorne J. 2009. A Software Architecture supporting Run-Time Configuration and Self-Management. *Communications of SIWN*. 7 (May. 2009), SIWN. 103-112.

[19] Pelc M., Anthony R. J., Ward P., and Hawthorne J. 2009. Practical Implementation of a Middleware and Software Component Architecture supporting Reconfigurability of Real-Time Embedded Systems. In *Proc. 7th IEEE/IFIP Intl. Conf. on Embedded and Ubiquitous Computing* (Vancouver, Canada, 2009). EUC'09. IEEE, 394-402.

A Metaheuristic Particle Swarm Optimization Approach to Nonlinear Model Predictive Control

Julian Mercieca and Simon G. Fabri
Department of Systems and Control Engineering
University of Malta
Msida MSD 2080, Malta
Email: julianmercieca@gmail.com, simon.fabri@um.edu.mt

***Abstract*—This paper commences with a short review on optimal control for nonlinear systems, emphasizing the Model Predictive approach for this purpose. It then describes the Particle Swarm Optimization algorithm and how it could be applied to nonlinear Model Predictive Control. On the basis of these principles, two novel control approaches are proposed and analysed. One is based on optimization of a numerically linearized perturbation model, whilst the other avoids the linearization step altogether. The controllers are evaluated by simulation of an inverted pendulum on a cart system. The results are compared with a numerical linearization technique exploiting conventional convex optimization methods instead of Particle Swarm Optimization. In both approaches, the proposed Swarm Optimization controllers exhibit superior performance. The methodology is then extended to input constrained nonlinear systems, offering a promising new paradigm for nonlinear optimal control design.**

Keywords-particle swarm optimization; model predictive control; optimal control; nonlinear control; computational intelligence; swarm intelligence; evolutionary intelligence; artificial intelligence; metaheuristic algorithms

I. INTRODUCTION

This paper discusses the use of Particle Swarm Optimization for optimal control of nonlinear systems. It proposes two novel control schemes in this regard and presents a more detailed perspective on earlier work by the same authors [1]. In the case of systems exhibiting linear dynamics, optimal and robust control theory offer well-developed tools to optimize a number of performance indices that embody desirable objectives and ensure performance robustness. These range from classical and fundamental robust control approaches [2] to more advanced, theoretically elegant and computationally tractable solutions [3], [4].

In contrast to linear optimal and robust control, its nonlinear counterpart (namely optimization constrained by a nonlinear dynamical system) is still a developing field. Its roots were laid down in the 1950s with the introduction of the Pontryagin maximum principle (a generalization of the Euler-Lagrange equations derived from the calculus of variations [5]) and dynamic programming, leading to the Hamilton-Jacobi-Bellman partial differential equations [6]. These were more theoretical contributions than practical design techniques. Numerous design methodologies have now been developed for nonlinear optimal control, often following different paths and techniques. The problem is attacked on several different fronts including extensions of linear theory, utilizing generalizations of the Lyapunov methodology, and brute force computation to name a few [7].

The advent of the microprocessor and the subsequent computer revolution opened up an entirely new possibility for optimal control: obtaining solutions directly through numerical computations. While the solution of the Hamilton-Jacobi-Bellman equation remains intractable in all but the simplest of cases, Euler-Lagrange type trajectory optimizations provide an alternative, more computationally feasible approach. Computers are able to provide relatively efficient solutions by solving trajectory optimizations that produce open-loop control trajectories as a function of time (as opposed to a state-feedback law). Feedback can then be incorporated by the repeated on-line solution of these trajectory optimizations, an approach known as receding or moving horizon. A heavy exploitation of the receding horizon methodology spawned the technique of Model Predictive Control [8]. Plants with slow dynamics were among the first candidate applications of this approach because on-line inter-sample computation of a sequence of manipulated variable adjustments in order to optimize the future behaviour of a controlled process using minimal control effort became feasible [9]. Additionally, the receding horizon strategy was a natural approach to constrained systems because constraints could be directly incorporated into the optimizations, enabling plant operators to run the plant near constraint boundaries, which can increase productivity and reduce product quality variation [10], [11]. Furthermore, Model Predictive Control seems extremely powerful for processes with dead-time or if the set-point is programmed. This is evidenced by its successful implementation in industrial process applications [8]–[13].

However, despite being an attractive control scheme for manipulating the behaviour of complex systems [14] and exhibiting excellent dynamic performance in both industrial applications and theoretical studies [15]–[17], the application of Model Predictive Control to nonlinear systems, known as Nonlinear Model Predictive Control (NMPC), is complicated largely due to the optimization method that has come to be used in these controllers. A fundamental difficulty of the NMPC approach is the requirement to solve nonconvex constrained optimization problems. Most existing works are based on nonlinear programming methods [18] that only

yield local optimum values, with the latter depending on the selection of the starting point. For this purpose, Alaniz [19] developed a particular numerical linearization technique to obtain a convex constrained optimization problem, albeit at the cost of performance deterioration. Other attempts to solve the nonconvex optimization problems exploit Genetic Algorithm (GA) optimizers [20]. However these face many challenges, including enormous computational effort due to its natural genetic operations [21], [22]. Although this may be reduced by using a real-value representation in the GA [21], [23], [24], some deficiencies in GA performance have been highlighted in recent research. Applications governed by highly epistatic objective functions [25], [26] reveal shortfalls in performance, which is further worsened by the GA's premature convergence [25].

This paper presents and analyses in depth two novel NMPC controllers based on a powerful optimization paradigm called Particle Swarm Optimization (PSO). PSO was first developed by Kennedy and Eberhart in 1995 [27]. This metaheuristic algorithm has been found to be robust in solving continuous nonlinear optimization problems [23], [26]–[28] and capable of generating high quality solutions with more stable convergence characteristics and shorter calculation times than other stochastic methods [23], [26], [29]. The salient feature of PSO lies in its learning mechanism, distinguishing it from other computational intelligence techniques, such as genetic algorithms, where PSO has been shown to have more attractive properties [30], [31]. PSO is governed by less tunable parameters and is notably easy to program and implement using basic logic and mathematical operations. The swarm intelligence algorithm stands in clear contrast with many optimization techniques for being derivative-free, being less sensitive to the objective function's nature, namely continuity and convexity, and not requiring good initial solutions for the iteration process to start. Furthermore, its flexibility enables its integration with other optimization techniques, forming hybrid tools [32]. Its ability to avoid local minima and to cater for stochastic objective functions, as is the case of representing a random optimization variable, further strengthens PSO's capacity to achieve superior optimization performance [32]. Owing to its simple concept and high efficiency, PSO has become a widely adopted optimization technique and has been successfully applied to many real-world problems [33]–[40]. Moreover, PSO's superiority is confirmed when compared with other optimization algorithms in various application areas [41]–[45]. Also, in the process of validating new global optimization techniques, researchers have proven that PSO performs well in several benchmark optimization problems [46]–[49].

One of the novel controllers presented in this paper is based on a numerical linearization technique first proposed by Alaniz in [19] that is based on conventional convex optimization methods. By contrast, the proposed controller exploits PSO techniques for optimization. The second novel controller proposed in this paper does away with any form of numerical linearization to achieve optimization of the cost function. Both controllers are simulated on an inverted pendulum on a cart problem and compared with the NMPC controller in [19].

The rest of the paper is organized as follows. Section II is an explanation of the implemented PSO algorithm, while Section III outlines the design of the three NMPC controllers evaluated in this paper. Section IV then presents the simulation setup, results and analysis, followed by a brief conclusion in Section V.

II. Particle Swarm Optimization

The particle swarm optimization algorithm is a population-based search algorithm inspired by the social behaviour of birds within a flock [27]. The very simple behaviour followed by individuals in a flock emulates their own successes and the success of neighbouring individuals. The emergent collective behaviour is that of discovering optimal regions of a high dimensional search space.

In a PSO algorithm, each particle representing a potential solution is maintained within a swarm. In simple terms, the particles are therefore "flown" through a multidimensional search space where the position of each particle is adjusted according to the experience of itself and its neighbours. Let $\mathbf{x}_i(t)$ denote the position of particle i in the search space at time step t, which denotes discrete time steps unless otherwise stated. The position of the particle is changed by adding a velocity vector, $\mathbf{v}_i(t)$, to the current position *i.e.,*

$$\mathbf{x}_i(t+1) = \mathbf{x}_i(t) + \mathbf{v}_i(t+1) \tag{1}$$

with $\mathbf{x}_i(0) \sim U(\mathbf{x}_{min}, \mathbf{x}_{max})$, where $U(\mathbf{x}_{min}, \mathbf{x}_{max})$ denotes the continuous uniform probability distribution within the real-valued space $(\mathbf{x}_{min}, \mathbf{x}_{max})$. The optimization process is driven by the velocity vector, reflecting both the experiential knowledge of the particle (known as *cognitive component*) and socially exchanged information from the particle's neighbourhood (known as *social component*). In this paper we implement a particular PSO algorithm know as global best PSO [50], which exhibits very fast convergence rates much needed for our predictive control application. For the global best PSO, or *gbest* PSO, the neighbourhood for each particle is the entire swarm, thus employing the social network of the star topology type. In this situation, the social information is the best position found by the swarm, referred to as $\hat{\mathbf{y}}(t)$.

For *gbest* PSO, the velocity of particle i is calculated as

$$\begin{aligned} v_{ij}(t+1) = \quad & v_{ij}(t) + c_1 r_{1j}(t)[y_{ij}(t) - x_{ij}(t)] + \\ & c_2 r_{2j}(t)[\hat{y}_j(t) - x_{ij}(t)] \end{aligned} \tag{2}$$

where $x_{ij}(t)$, $y_{ij}(t)$ and $v_{ij}(t)$ are the position, personal best position and velocity of particle i in dimension $j = 1, \ldots, n_x$ at time step t respectively, $\hat{y}_j(t)$ is the global best position in dimension j, c_1 and c_2 are positive acceleration constants used to scale the contribution of the cognitive and social components respectively, and $r_{1j}(t), r_{2j}(t) \sim U(0,1)$ are random values in the range $[0,1]$, sampled from a continuous uniform distribution. These random values introduce a random element to the algorithm.

The personal best position, $\mathbf{y}_i$, associated with particle i is the best position the particle has visited since the first time

step. Considering a minimization problem, the personal best position at the next time step, $t+1$, is calculated as

$$\mathbf{y}_i(t+1) = \begin{cases} \mathbf{y}_i(t) & \text{if } f(\mathbf{x}_i(t+1)) \geq f(\mathbf{y}_i(t)) \\ \mathbf{x}_i(t+1) & \text{if } f(\mathbf{x}_i(t+1)) < f(\mathbf{y}_i(t)) \end{cases} \quad (3)$$

where $f : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ is the fitness function, which is a measure of how close the corresponding solution is to the optimum, quantifying the performance, or quality, of a particle (or solution).

The global best position, $\hat{\mathbf{y}}(t)$, at time step t, is defined as

$$\hat{\mathbf{y}}(t) \in \{\mathbf{y}_0(t) \dots \mathbf{y}_{n_s}(t)\} | f(\hat{\mathbf{y}}(t)) = \min\{f(\mathbf{y}_0(t)) \dots f(\mathbf{y}_{n_s}(t))\} \quad (4)$$

where n_s is the total number of particles in the swarm. Equation 4 therefore states that $\hat{\mathbf{y}}(t)$ is the best position discovered by any of the particles so far. In this paper, it is calculated as the best personal-best position. Algorithm 1 summarizes the *gbest* PSO algorithm.

Further to the basic PSO algorithm just described, the speed of convergence and quality of solutions are improved using velocity clamping and inertia weight [51]. The efficiency and accuracy of our optimization algorithm is governed by the exploration-exploitation trade-off [52]. *Exploitation* is the ability of a search algorithm to concentrate the search around a promising area in order to refine a candidate solution. *Exploration*, on the other hand, is the ability to locate a global optimum by exploring different regions of the search space. A good optimization algorithm balances these contradictory objectives in an optimal manner. For the PSO algorithm, these objectives are reached using the velocity update equations.

Algorithm 1 *gbest* PSO [50]

Create and initialize an n_x-dimensional swarm
repeat
  for *each particle* $i = 1, \dots, n_s$ **do**
    //set the personal best position
    if $f(\mathbf{x}_i) < f(\mathbf{y}_i)$ **then**
      $\mathbf{y}_i = \mathbf{x}_i$;
    end
    //set the global best position
    if $f(\mathbf{y}_i) < f(\hat{\mathbf{y}})$ **then**
      $\hat{\mathbf{y}} = \mathbf{y}_i$;
    end
  end
  for *each particle* $i = 1, \dots, n_s$ **do**
    update the velocity using Equation (2);
    update the position using Equation (1);
  end
until *stopping condition is true*;

The velocity update presented in Equation (2) comprises three terms that contribute to the step size of particles. The early applications of basic PSO revealed that the velocity quickly explodes to large values, especially for particles located far from the neighbourhood best and personal best positions. As a consequence, particles have large position updates resulting in them leaving the boundaries of the search space and diverging. The global exploration of particles may be controlled by clamping velocities to stay within boundary constraints [53]. If a specified maximum velocity is exceeded, the particle's velocity is set to the maximum velocity. Let $V_{max,j}$ denote the maximum allowed velocity in dimension j. Particle velocity is then adjusted prior to the position update using,

$$v_{ij}(t+1) = \begin{cases} v_{ij}'(t+1) & \text{if } v_{ij}'(t+1) < V_{max,j} \\ V_{max,j} & \text{if } v_{ij}'(t+1) \geq V_{max,j} \end{cases} \quad (5)$$

where v_{ij}' is calculated using Equation (2).

The value of $V_{max,j}$ is essential to control the granularity of the search by clamping escalating velocities. Large values of $V_{max,j}$ facilitate global exploration, while smaller values encourage local exploitation. Too small values of $V_{max,j}$ leads to insufficient exploration beyond locally good regions, increasing the number of time steps to reach an optimum, with the risk of the swarm becoming trapped in a local optimum, with no means of escape. On the other hand, too large values of $V_{max,j}$ risk the possibility of missing a good region, having the particles possibly jumping over good solutions and continuing to search in fruitless regions of the search space. Despite this disadvantage of particles possibly jumping over optima, particles move faster.

The problem of finding a good value for each $V_{max,j}$ still stands. We require the balance between (a) moving too fast or too slow, and (b) exploration and exploitation. Here, we select $V_{max,j}$ values to be a fraction of the domain of each dimension of the search space. That is,

$$V_{max,j} = \delta(x_{max,j} - x_{min,j}) \quad (6)$$

where $x_{max,j}$ and $x_{min,j}$ are the maximum and minimum values of the domain of $\mathbf{x}$ in dimension j respectively, and $\delta \in (0,1]$. In a number of empirical studies it was found that the value of δ is problem-dependent [54], [55], and the best value for our situation was therefore obtained empirically.

While velocity clamping exhibits the advantage of a controlled explosion of velocity, it also presents a difficulty when all velocities are equal to the maximum velocity. If no precautionary measures are implemented, particles remain searching on the boundaries of a hypercube defined by $[\mathbf{x}_i(t) - \mathbf{V}_{max}, \mathbf{x}_i(t) + \mathbf{V}_{max}]$. A particle may stumble upon the optimum, but in general exploiting this local area is difficult. This problem is solved in our algorithm by introducing an inertia weight, a concept introduced by Eberhart and Shi [28] as a mechanism of controlling the exploitation and exploration abilities of the swarm, with the original intention of eliminating the need for velocity clamping [33]. The inertia weight was successful in addressing the first objective, but failed to completely eliminate the need for velocity clamping. The inertia weight, w, controls the particle's momentum by weighting the contribution of the previous velocity - in other words, controlling how much memory of the previous flight

direction will influence the new velocity. For the *gbest* PSO, the velocity equation changes from Equation (2) to

$$v_{ij}(t+1) = wv_{ij}(t) + c_1 r_{1j}(t)[y_{ij}(t) - x_{ij}(t)] + c_2 r_{2j}(t)[\hat{y}_j(t) - x_{ij}(t)] \quad (7)$$

The value of w is essential to ensure convergent behaviour while optimally trading off exploration and exploitation. For $w \geq 1$, velocities increase over time, accelerating towards the maximum velocity (assuming a velocity clamping strategy), and the swarm diverges. Particles fail to change direction to return back towards promising areas. For $w < 1$, particles undergo a deceleration until their velocities reach zero (depending on the values of the acceleration coefficients). Large values of w facilitate exploration with increased diversity, while small w promotes local exploitation. However, very small values eliminate the exploration ability of the swarm since little momentum is then preserved from the previous time step, enabling quick changes in direction. The smaller w, the more do the social and cognitive components control position updates.

As with the maximum velocity, the optimal value for the inertia weight is problem-dependent [55]. Here, we make use of dynamically changing inertia values, starting with large inertia values that decrease over time to smaller values. This way, particles are allowed to explore in the initial search steps, while favouring exploitation as time increases.

The inertia weight is dyamically varied using the linear decreasing method, where an initially large inertia weight (usually 0.9) is linearly decreased to a small value (usually 0.4). Following Yoshida *et al.* [56], Suganthan [57], Ratnaweera *et al.* [58], Naka *et al.* [59], we set

$$w(t) = (w(0) - w(n_t))\frac{(n_t - t)}{n_t} + w(n_t) \quad (8)$$

where n_t is the maximum number of time steps that the algorithm is executed, $w(0)$ is the initial inertia weight, $w(n_t)$ is the final inertia weight, and $w(t)$ is the inertia at time step t. Note that $w(0) > w(n_t)$.

III. Nonlinear Model Predictive Control

A nonlinear dynamic system may be represented by a set of nonlinear differential equations [60] that may be discretized for computational purposes using Euler's method, where T_s is the sampling period and k is the sample index in discrete-time, as follows:

$$\mathbf{x}(k+1) = \mathbf{x}(k) + T_s f(\mathbf{x}(k), \mathbf{u}(k), \mathbf{v}(k), k) \quad (9)$$

$$\mathbf{y}(k) = g(\mathbf{x}(k), \mathbf{u}(k), \mathbf{v}(k), k) \quad (10)$$

Arguments of the nonlinear function f include a state vector $\mathbf{x}(k)$, a control input $\mathbf{u}(k)$, and a disturbance input $\mathbf{v}(k)$. The set of physical quantities that can be measured from the system constitute the output, $\mathbf{y}(k)$, which is also a nonlinear function g of the same arguments. More accurate discretization approximations, such as the Runge-Kutta methods, can be used if the system dynamics are highly nonlinear or the sampling period is large.

The Model Predictive Control (MPC) design methodology is characterized by three main features: an explicit model of the plant, computation of control signals by optimizing predicted plant behaviour, and a receding horizon [10]. MPC's receding horizon strategy can be explained using Figure 1. An internal model predicts how the plant will react, starting at the current time k, over a discretized prediction interval. The letter l denotes the number of discrete steps in this interval. Each discrete step spans a time of T_s seconds, therefore the prediction interval lasts lT_s seconds. The predicted behaviour is governed by the present state $\mathbf{x}(k)$, an estimated disturbance history $\mathbf{v}$, and a control history $\mathbf{u}$ that is to be applied. The objective is to select the control history that yields the best predicted behaviour with respect to a reference trajectory and optimization parameters.

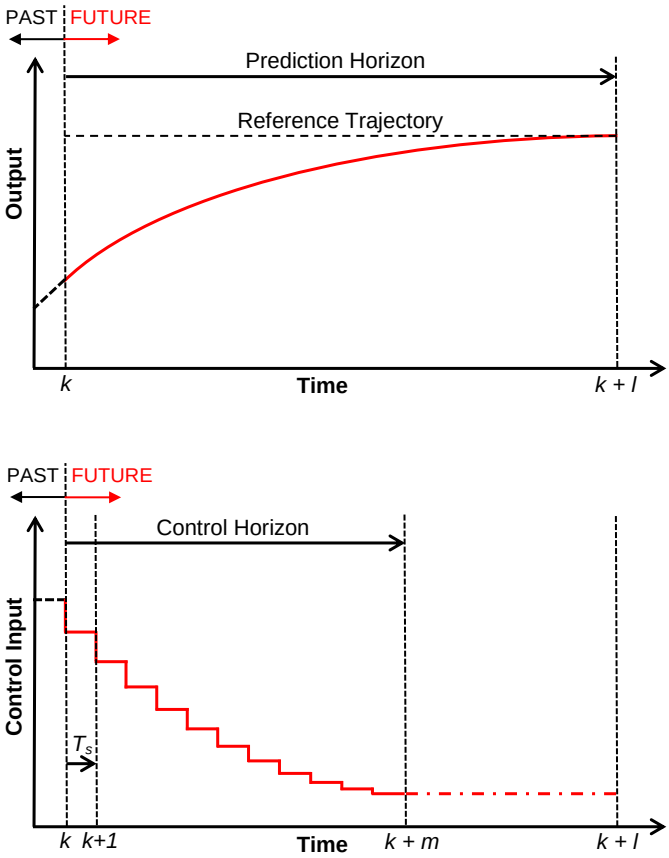


Fig. 1. Nonlinear Model Predictive Control: A receding horizon strategy

MPC solves for the control history, which is a sequence of m vector values. Two adjacent control values are separated by a time step of T_s, therefore having a control history spanning mT_S seconds. During each time step, the control values are held constant and the values are assumed to change instantaneously as soon as a new time step is started. After the control history has ended, the control signal is held constant until the prediction interval is over.

Once the optimal control history has been computed, the first N time steps of the solution are applied to the plant and

the rest are discarded. After these N time steps have passed, the cycle of forming predicted behaviours and computing the control history is repeated. In this paper, we choose N to be unity, however this number can be increased to reduce the rate of production of solutions.

The cost function used in our performance evaluation is given by Equation (11) having a quadratic structure comprising two terms. The first term, weighted by a symmetric weighting matrix $\mathbf{Q}(k)$, penalizes the deviations from a reference trajectory that occur throughout the prediction interval. A specific value of the reference is denoted by $\tilde{\mathbf{y}}(k)$. The second term, weighted by a symmetric weighting matrix $\mathbf{R}(k)$, penalizes the magnitude of each control value in the control history.

$$J = \sum_{i=0}^{l-1} ||(\mathbf{y}(k+i) - \tilde{\mathbf{y}}(k+i))||^2_{\mathbf{Q}(k+i)} + \sum_{i=0}^{m-1} ||\mathbf{u}(k+i)||^2_{\mathbf{R}(k+i)} \tag{11}$$

As previously described, if m is in the range $1 \leq m \leq l$, then the last value in the control history is held constant for the final $(l-m)$ time steps. The value of $\mathbf{R}(k+m-1)$ should therefore have a different magnitude to compensate for the added duration of $\mathbf{u}(k+m-1)$.

Control and output constraints are considered in their inequality form, with the constraints being enforced at each discretized point in the control history and output trajectory, as shown in equations (12) and (13).

$$\mathbf{u}(k)_{min} \leq \mathbf{u}(k) \leq \mathbf{u}(k)_{max} \tag{12}$$

$$\mathbf{y}(k)_{min} \leq \mathbf{y}(k) \leq \mathbf{y}(k)_{max} \tag{13}$$

The MPC problem in this setting is to minimize J by choosing $\mathbf{u}$, subject to the constraints in equations (12) and (13) and the dynamics of equations (9) and (10). We will now describe the three nonlinear model predictive controllers considered in this paper, two of them representing the novel contributions of this work.

A. A numerical linearization method

This method, proposed by Alaniz in [19], centres around a particular numerical linearization technique for generating the predicted output trajectory $\mathbf{y}$. A nominal control history $\bar{\mathbf{u}}$ is first chosen, then the corresponding nominal output trajectory $\bar{\mathbf{y}}$ is computed through numerical integration. Typically $\bar{\mathbf{u}}$ is the previous optimal solution, but it can be set equal to zero if none exist. The predicted output is then based on linearizing the control perturbation $\Delta\mathbf{u}$ about the nominal trajectory as follows:

$$\begin{aligned}
\mathbf{y}(k) &= \bar{\mathbf{y}}(k) + \boldsymbol{\alpha}_0 \Delta\mathbf{u}(k) \\
\mathbf{y}(k+1) &= \bar{\mathbf{y}}(k+1) + \boldsymbol{\alpha}_1 \Delta\mathbf{u}(k) + \boldsymbol{\beta}_0 \Delta\mathbf{u}(k+1) \\
\mathbf{y}(k+2) &= \bar{\mathbf{y}}(k+2) + \boldsymbol{\alpha}_2 \Delta\mathbf{u}(k) + \boldsymbol{\beta}_1 \Delta\mathbf{u}(k+1) + \\
&\quad \boldsymbol{\gamma}_0 \Delta\mathbf{u}(k+2) \\
&\vdots
\end{aligned} \tag{14}$$

The coefficients $\boldsymbol{\alpha}_i$, $\boldsymbol{\beta}_i$, $\boldsymbol{\gamma}_i$, ... are produced by computing a perturbed trajectory for each $\Delta\mathbf{u}(k+i)$ and finding the subsequent deviation from the nominal trajectory. Perturbed trajectories are the result of adding a pulse of magnitude one to the nominal control history at time $(k+i)$. Each trajectory is formed by propagating the present state $\mathbf{x}(k)$ over a fixed interval of time while applying an associated control history. The prediction interval and control history are divided into l and m discrete steps, respectively, of length T_s where $m \leq l$. After the control history has ended, it is held constant for the final $(l-m)$ time steps.

The MPC problem is to solve for the optimal control perturbation $\Delta\mathbf{u}^*$ by minimizing a cost function with respect to a reference trajectory and optimization parameters. The optimal control history is then the sum of the nominal control history and the optimal control perturbation [19]. By rearranging and simplifying the form of Equation (11), a set of matrices is obtained that leads to the unconstrained and constrained optimization problems. For the unconstrained case, Alaniz [19] presents a solution by using an equivalent least squares technique, while for the constrained case, the problem is reinterpreted so as to obtain the standard form handled by quadratic programming solvers.

Once the optimal control history is chosen, the first N time steps of the solution are applied to the plant. The cycle of forming predicted behaviours and solving for the optimal control perturbations is then repeated using the most recent feedback from the plant. The interested reader is referred to [19] for further detail about this technique.

B. A novel numerical linearization technique using PSO

A novel application of PSO proposed here exploits the aforementioned numerical linearization technique used in conjunction with the PSO algorithm, where the convex least squares or quadratic programming optimization methods are now replaced by the global best PSO algorithm. The evaluation function is the cost given by Equation (11), so that PSO searches for the optimal perturbed control history of Equation (14), denoted by $\Delta\mathbf{u}(k)^*$, in order to obtain the optimal control history $\mathbf{u}(k)^*$ that minimizes J. For this purpose we require an m-dimensional PSO, with each particle's position defined by $\mathbf{K}$, an m-dimension column vector equal to $\Delta\mathbf{U}(k)^*$, which is a column vector having $\Delta\mathbf{u}(k+i)^*$ as its elements.

C. A novel PSO-based nonlinear MPC strategy

The second novel controller makes use of the PSO search algorithm for obtaining the optimal control history that minimizes directly the cost function J given by Equation (11) without resorting to numerical linearization as represented by Equation (14). In this manner we simply use Equation (11) as the evaluation function to be minimized using global best PSO, thereby avoiding any linearization technique or mathematical result for minimization, albeit at an increased computational complexity. Each particle's position in the swarm represents the m-dimension column vector defining the optimal control history, $\mathbf{U}(k)^*$.

As we shall see, this remarkably straightforward approach produces the best results for the controllers studied in this

paper in terms of the performance index obtained. The block diagram in Figure 2 illustrates the structure of the proposed predictive control loop. A particle swarm optimizer uses the reference input and predicted output trajectories to minimize the quadratic cost function given by Equation (11) and compute the optimal control history that is then applied to the plant. The proposed controller is further enhanced by actively correcting the weighting matrix **R** in an adaptive manner, so that the chattering effect of the control input observed about the equilibrium point is reduced.

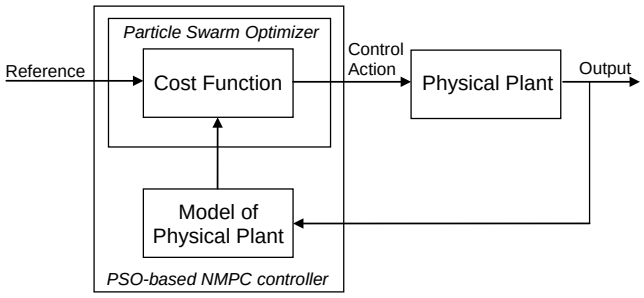


Fig. 2. PSO-based nonlinear MPC loop

IV. Performance Evaluation: Inverted Pendulum on Cart

The performance of the proposed controllers is evaluated by analyzing the results from simulation experiments. The plant chosen for simulation is an inverted pendulum on a cart and two types of controllers are generated for the three methods presented in the previous section; an unconstrained and constrained version. The latter problem shall only consider a single constraint that restricts the input as per the inequality given by Equation (12). Hence, no penalty functions are required. The pendulum is initially at the stable equilibrium point and the purpose of each controller is to invert the pendulum. Since the dynamics at the stable and unstable equilibrium points are very different, this is a good problem to demonstrate the effectiveness of our nonlinear MPC controllers.

A. Plant Model

The nonlinear model of the plant is derived by applying Newton's Laws of Motion to the free body diagrams in Figure 3. The resulting equations of motion are given by equations (15) and (16). A complete derivation is given in [19].

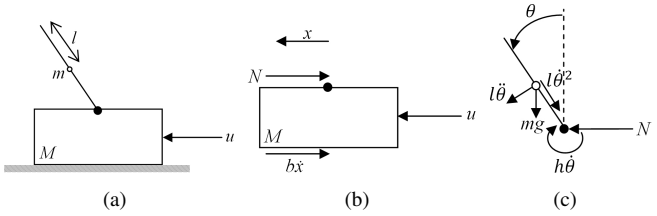


Fig. 3. (a) Inverted Pendulum on a Cart; (b) Free body diagram 1 (cart); (c) Free body diagram 2 (pendulum).

$$\ddot{x} = \frac{1}{M+m}[u - b\dot{x} - ml\ddot{\theta}\cos(\theta) + ml\dot{\theta}^2\sin(\theta)] \quad (15)$$

$$\ddot{\theta} = \frac{3}{4ml^2}[mgl\sin(\theta) - ml\ddot{x}\cos(\theta) - h\dot{\theta}] \quad (16)$$

M represents the mass of the cart that slides along a surface, m is the uniformly distributed mass of an ideal pendulum, $2l$ is the length of the ideal pendulum, b is the surface friction damping constant, h is the rotational friction damping coefficient, u is the force applied to the block, θ is the clockwise angle between the normal and the pendulum (as shown in Figure 3(c)), and x is the cart's horizontal displacement from its equilibrium position. To allow the model to be numerically integrated, equations (15) and (16) are expressed in terms of the state variables x, $\dot{x}$, θ, and $\dot{\theta}$. The second-order differential equations have the form given by Equation (17), where $\boldsymbol{\chi}$ is a vector variable. A vector field g is also created to combine the states into one state vector. The second-order differential equations are then discretized using the fourth-order Runge-Kutta method [19].

$$\ddot{\boldsymbol{\chi}} = f(\dot{\boldsymbol{\chi}}, \boldsymbol{\chi}, u), \quad \begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix} = g(\dot{\boldsymbol{\chi}}, \boldsymbol{\chi}), \; \boldsymbol{\chi} = \begin{bmatrix} x \\ \theta \end{bmatrix} \quad (17)$$

B. Controller Layout

The simulation experiments were run on the Simulink software package [61]. The layout shown in Figure 4 is the simulated realization of the control loop given in Figure 2. It makes use of Matlab S-Functions that implement constrained or unconstrained versions of the PSO-based NMPC controller.

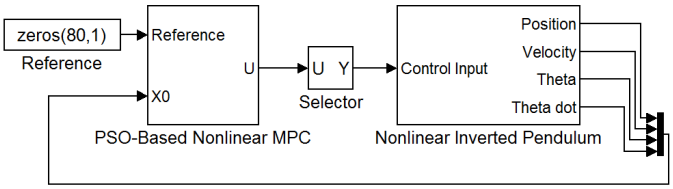


Fig. 4. Nonlinear MPC Simulink layout

C. Controller Parameters

The MPC controller rate is $\frac{1}{NT_s}$, where N is the number of controls in the control history that are applied to the plant. $N = 1$ is used in the controller since this is the typical value selected in MPC [10]. The computational load of MPC can be reduced if N is increased, but a disadvantage to having $N > 1$ is that some of the controls applied to the plant are based on old feedback. The fourth-order Runge-Kutta method is tested using different values for T_s, and it is established that the response with $T_s = 0.1s$ is almost indistinguishable from the actual response, thus using this value for the controller.

Since this controller is very computationally intensive, it is not feasible to have a long prediction length or control history. A value of $l = m = 20$ is chosen as a balance between

performance and computation time. This results in a controller capable of predicting for 2 seconds.

The two novel PSO-based controllers use the following PSO parameters, which were derived empirically through successive simulations:

- Each particle consists of 20 members, corresponding to the 20 elements that make up the optimal control perturbation history column vector, $\Delta\mathbf{U}(k)^*$, for the PSO-based numerical linearization method, or the optimal control history column vector, $\mathbf{U}(k)^*$, for the PSO-based NMPC controller.
- Swarm size, n_s = 30.
- Inertia weight w is set by Equation (8), where $w(0) = 0.9$ and $w(n_t) = 0.4$.
- Velocity clamping is governed by equations (5) and (6), with $\delta = 0.5$.
- Search space is limited to real-valued variables between $-300N$ ($x_{min,j}$) and $300N$ ($x_{max,j}$) for the unconstrained case.
- Acceleration coefficients $c_1 = 2$ and $c_2 = 2$.
- Number of iterations = 30.

Both weighting matrices $\mathbf{Q}$ and $\mathbf{R}$ given in Equation (11) are set equal to the values shown in Equation (18), which includes a vector showing the order of the outputs. The objective of this controller is to invert the pendulum. Hence the penalty on θ is increased relative to the other states so that the controller focuses more on decreasing angular deviations than other state deviations. Initially, the cart must move back and forth until the pendulum gains enough momentum to swing up. Increasing the θ penalty would reduce the time required for the pendulum to swing up. Through successive simulation trials, a value of 100 was finally chosen as the penalty on the θ state by comparing the time required to achieve pendulum inversion.

$$\mathbf{y} = \begin{bmatrix} x \\ \dot{x} \\ \theta \\ \dot{\theta} \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 100 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \ R = 1 \qquad (18)$$

Deviations are measured from the reference trajectory that is set equal to a zero column vector, as given by Equation (19). In this example, we set each reference state variable to zero at each time step of the prediction interval. This reference remains constant for the duration of the simulation.

$$\tilde{\mathbf{Y}} = \begin{bmatrix} 0 & 0 & \cdots & 0 \end{bmatrix}^T \qquad (19)$$

D. Simulation Results

The responses of the unconstrained controller, using the three control schemes described in sections IIIA, IIIB, and IIIC, are shown in Figure 5 as blue, green and red curves respectively. The pendulum is initially set at the stable equilibrium point (180 degrees), hanging straight down. The performance index graph plots the value of the cost function J for each time step as calculated by the control law using Equation (11). The cart's position moves back and forth so that the pendulum gains momentum. This continues until there is enough momentum to swing up and invert the pendulum in the 0 degree position. Table I shows typical performance results obtained in this unconstrained control case. The performance index values quoted in the table are obtained from Equation (11) but this time using the actual output trajectory, instead of the predicted one, and the actual control inputs applied, instead of the computed control history. The summation is calculated for a sufficiently large amount of time ensuring that the system has settled into steady state. Table I therefore reveals the *true* performance index for the whole control action in the unconstrained case, demonstrating the typically superior performance of the proposed PSO-based NMPC controller. The results show that when PSO is used in conjunction with the numerical linearization technique, only a minimal advantage is obtained over the least squares method (an improvement in J of only 1.46%), as expected for the convex optimization problem being solved. On the other hand, the second proposed nonlinear PSO controller gives a significant improvement in J of 8.04% over the numerical linearization (least squares) counterpart.

TABLE I
UNCONSTRAINED NMPC: TRUE PERFORMANCE INDEX VALUES

Method	Numerical Linearization (Least Squares) [19]	Numerical Linearization (PSO)	PSO
True Performance Index J ($\times 10^6$)	1.4538	1.4326	**1.3369**

Figure 6 shows the response plots of the constrained controller when a single constraint, restricting the control input of the cart to be within $-45N$ and $45N$, is made active. In Figure 6, the final angular deflection is either 0° or 360°. Note that both these angles correspond to the same inverted position of the pendulum. For the novel PSO-based NMPC controller, the cart is noted to move a much smaller distance to achieve swing-up. In real-world terms, this translates to a more efficient process, with less work being done by the cart to achieve swing-up and equilibrium. This is further evidenced by Table II, which indicates that the novel PSO-based nonlinear MPC controller has the edge over the numerical linearization technique that uses quadratic programming, a method known to have problems in getting stuck at local minima [62]. We record a 14.78% improvement in J, accompanied by a very low standard deviation when the experiment is repeated over

TABLE II
CONSTRAINED NONLINEAR MPC: TRUE PERFORMANCE INDEX VALUES

Method	Numerical Linearization (Quadratic Programming) [19]	PSO (mean J)	PSO (standard deviation) ($\times 10^6$)
True Performance Index J ($\times 10^6$)	2.8534	**2.4316**	**0.02396**

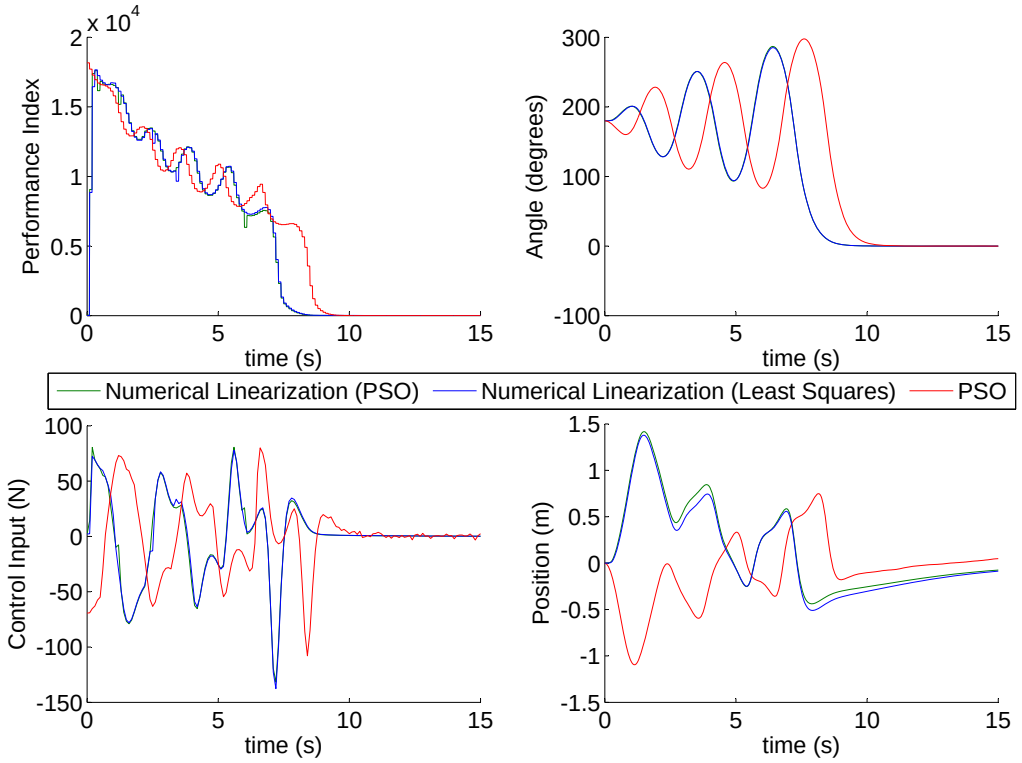


Fig. 5. Unconstrained nonlinear MPC: A comparison

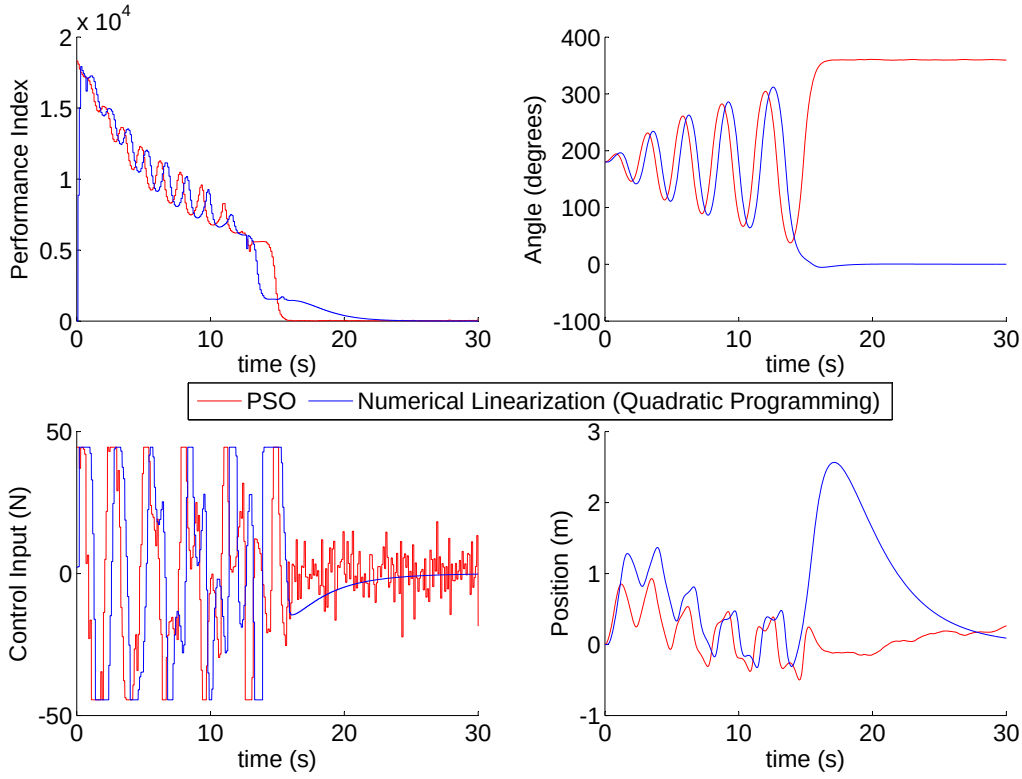


Fig. 6. Constrained nonlinear MPC: Restricting control input to within $-45N$ and $45N$ (10 independent trials)

10 trials indicating PSO's repeatable nature despite being a metaheuristic optimization method. Note that for the constrained case, using the numerical linearization technique in conjunction with PSO is computationally inefficient since every particle must be checked for its corresponding optimal control history, doubling the workload of its unconstrained counterpart, rendering it practically useless to investigate for this purpose.

The advantage of the novel PSO-based NMPC controller is even more evident in Figure 7, where the proposed active correction for the chattering effect of the control input is implemented for the same constrained NMPC problem by dynamically changing **R** appearing in Equation (11). The control input is being more heavily penalized when the angle approaches the equilibrium point by increasing R from 1 to 30. In other words, we are telling the system that in the close neighbourhood of the equilibrium point, minimal control effort is required, mitigating the effect of metaheuristic stochasticity. This reduces the true performance index even further, giving an improvement in J of 16.65% (see Table III), making the process even more efficient.

TABLE III
TRUE PERFORMANCE INDEX VALUE COMPARISON FOR ACTIVE R CORRECTION

Method	Numerical Linearization (Quadratic Programming) [19]	PSO
True Performance Index J ($\times 10^6$)	2.8534	**2.3810**

The system's robustness to model uncertainty is best illustrated by the simulation results of Figure 8. This is tested by randomly increasing or decreasing each of the plant model's parameters by 5% (all parameters are changed for every trial). Thus, the constrained NMPC controller is using a severely inaccurate model for its predictions and we are not actively controlling the control input weight R (to consider the worst case). Despite these adverse conditions, the results of Figure 8 show excellent performance and the pendulum swings up normally except for a larger distance now required. Table IV shows the corresponding changes implemented in the model parameters of one particular trial picked up at random in the simulation results of Figure 8. The corresponding performance index values are given in Table V, where although both controllers manage swing-up and equilibrium similarly as for the results shown in Figure 6, the novel PSO-based NMPC controller exhibits an improvement in J of 12.07%.

Repeatability is tested by performing several trials with different constraints, as shown in Table VI. The novel PSO nonlinear controller shows consistently better performance, with a mean improvement in J of 8.73%.

E. Scalability and Constraints

The inverted pendulum considered in these simulation experiments is a non-trivial example of a control engineering problem. It is characterized by nonlinear dynamics of an

TABLE IV
ACTUAL PLANT AND MODEL PARAMETERS (FOR A PARTICULAR TRIAL)

Parameter	Units	Actual Plant	Model
M	Kg	14.6	15.33
m	Kg	7.3	6.935
$2l$	m	2.4	2.52
b	Kg/s	14.6	15.33
h	Kgm^2/s	0.0136	0.0129

TABLE V
TRUE PERFORMANCE INDEX VALUES OBTAINED FOR THE ROBUSTNESS TEST

Method	Numerical Linearization (Quadratic Programming) [19]	PSO
True Performance Index J ($\times 10^6$)	2.7369	**2.4065**

TABLE VI
SIMULATION RESULTS FOR DIFFERENT CONSTRAINTS (10 INDEPENDENT TRIALS WITH CONSTANT R)

Constraint	Numerical Linearization (Quadratic Programming) [19]	PSO
$-30N \leq U \leq 30N$	2.7182	**2.4026**
$-35N \leq U \leq 35N$	2.5374	**2.3947**
$-40N \leq U \leq 40N$	2.4590	**2.3880**
$-45N \leq U \leq 45N$	2.8534	**2.4316**

overall order of four, dynamic interaction between the state variables, and under-actuation (one control input and two controlled outputs). Nevertheless, there do exist more complex plants having higher order and/or more inputs and outputs which would require a scale up of the proposed PSO-based controllers to higher dimensions. Current literature indicates that the PSO's performance remains robust even when applied to relatively high dimensions as typically found in control applications (in the order of tens). Indicative of this is Engelbrecht's work [63], which reveals the *gbest* PSO algorithm's superior performance with respect to all other homogeneous PSO algorithms considered therein. This is investigated for several benchmark unimodal and multimodal functions, and the issue of scalability for the *gbest* PSO algorithm starts becoming increasingly pronounced only for dimensions of order hundred or above. Following this empirical analysis, the heterogeneous PSO algorithm proposed by Engelbrecht in [64] is shown to be significantly more scalable to higher dimensions when compared with other algorithms [63], [65]. The foregoing analysis is further confirmed by Piccand *et al.* [66], who show how the *gbest* PSO algorithm employed in this paper exhibits a unity success rate for all the benchmark unimodal and multimodal functions considered therein, up to several tens of dimensions. Increasing the swarm size may sometimes be necessary to handle higher dimensions, however such tuning is also problem dependent [66]. In view of this published evidence, we envisage that for systems governed by increased state variables and/or inputs and outputs, the proposed PSO controllers are not expected to perform less

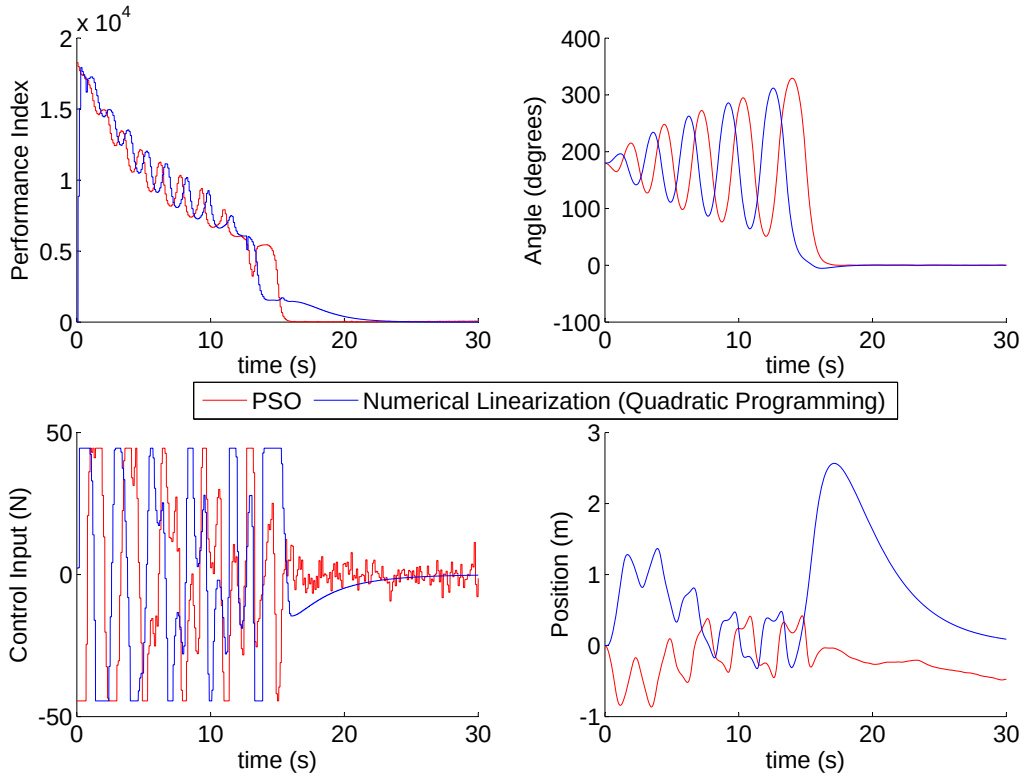


Fig. 7. Actively controlling control input weight R for reduced chattering.

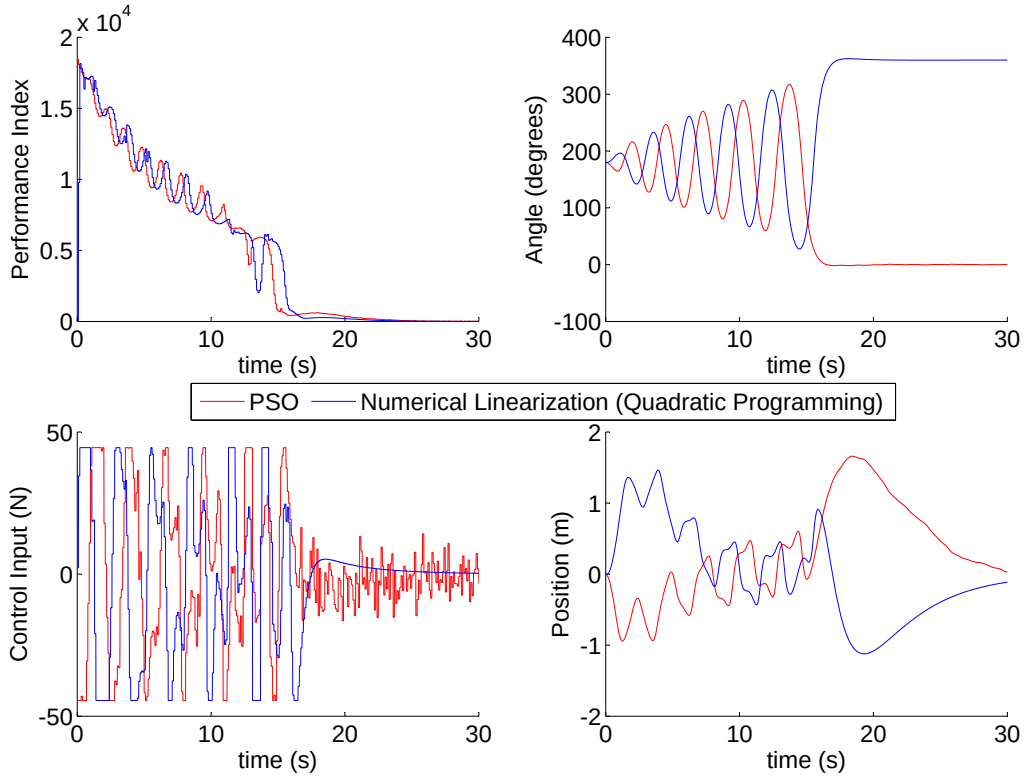


Fig. 8. Robustness Test: Model's parameters are significantly different from the actual plant parameters (constrained NMPC problem results shown).

reliably than the presented example. The process will necessarily be more computationally expensive, however this is an implementation issue which goes beyond the current scope of this work.

This paper has only considered the control input constrained problem for the NMPC case. Implementing PSO-based controllers for both control and output constraints, as given by equations (12) and (13), would require an elegant constraint handling approach. As with most applications of PSO, this could include the use of penalty methods to penalize those particles that violate constraints [67]–[69]. Other solutions, as reported by Shi and Krohling [70] and Laskari *et al.* [71], convert the constrained problem to an unconstrained Lagrangian. Also, repair methods, which allow particles to roam into feasible space, may be implemented by applying repairing operators to change infeasible particles to represent feasible solutions. A few examples include the work of Hu and Eberhart [72] that developed an approach where particles are not allowed to be attracted by infeasible particles, and that of El-Gallad *et al.* [73], which replaced infeasible particles with their feasible personal best positions. Other works by Venter and Sobieszczanski-Sobieski [74], [75] proposed a way of repairing infeasible solutions. Such methods may be investigated to determine the most efficient way of dealing with both input and output constraints.

V. Conclusion and Future Work

This paper has addressed the use of PSO for the design of model predictive control as applied to nonlinear systems. Following a detailed description of PSO and MPC theory, two novel controllers were proposed for the receding horizon model predictive strategy when applied to nonlinear dynamic systems. Both controllers exploit the well-known desirable properties of PSO. One makes use of a numerical linearization technique where, instead of convex optimization methods, we employed a PSO strategy. As expected, when simulated on an inverted pendulum on cart problem, this technique only yielded a minor improvement in performance over its convex optimization counterpart. By contrast, the second proposed controller proved superior to both, approaching up to 16% less performance cost at best. In addition, we proposed a further enhancement for this novel scheme by actively controlling the control input weight $\mathbf{R}$ to reduce the chattering effect of the control signal that is often observed in nonlinear model predictive control. This framework was also shown to be extensible to input constrained systems, thereby providing a foundation to include other advances in control theory as they become available.

This work may be further extended by an investigation on the use of PSO to obtain the much needed connection between the selection of $\mathbf{Q}$ and $\mathbf{R}$ (the two weighting matrices) and the performance specifications; possibly through some time-domain performance criterion. A similar investigation may be carried out for other control schemes, including linear quadratic optimal control strategies. Another interesting idea for future work is to run two optimizers in parallel. One strategy could employ a random search algorithm [76]–[78] running in parallel with PSO. Their possible collaboration could yield the right answer faster, whilst also being more robust to pathological cases. Furthermore, different variations of the PSO algorithm could be implemented, comparing their performance in the process, and also addressing scalability issues.

Having successfully implemented PSO for an NMPC problem, one should note that in certain circumstances when the problem is very well known, gradient-based methods may solve the NMPC problem more efficiently than PSO, albeit not as accurately. This calls for a detailed comparison between a PSO-based implementation and, for instance, one using a state-of-the-art NMPC solver such as the SQP-based method in ACADO [79], [80], or ICLOCS in combination with the interior point solver IPOPT [81]–[83].

All these issues may be investigated in future work for the purpose of establishing more effective ways of optimizing the NMPC problem.

References

[1] J. Mercieca and S. G. Fabri, "Particle swarm optimization for nonlinear model predictive control," in *Proceedings of the Fifth International Conference on Advanced Engineering Computing and Applications in Science - ADVCOMP 2011*, Lisbon, Portugal, November 2011, pp. 88–93.

[2] B. Anderson and J. Moore, *Optimal control: linear quadratic methods*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1990.

[3] M. Dahleh and J. Pearson, J., "I-optimal feedback controllers for mimo discrete-time systems," *IEEE Transactions on Automatic Control*, vol. 32, no. 4, pp. 314–322, Apr 1987.

[4] J. Doyle, K. Glover, P. Khargonekar, and B. Francis, "State-space solutions to standard H_2 and H_∞ control problems," *IEEE Transactions on Automatic Control*, vol. 34, no. 8, pp. 831–847, Aug 1989.

[5] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. Mishchenko, *The mathematical theory of optimal processes (International series of monographs in pure and applied mathematics)*. Interscience Publishers, 1962.

[6] R. Bellman, "On the Theory of Dynamic Programming," in *Proceedings of the National Academy of Sciences*, vol. 38, 1952, pp. 716–719.

[7] L. D. Berkovitz and N. G. Medhin, *Nonlinear Optimal Control Theory*, ser. Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, 2012.

[8] C. E. Garcia, D. M. Prett, and M. Morari, "Model predictive control: theory and practice - a survey," *Automatica*, vol. 25, pp. 335–348, May 1989.

[9] J. Richalet, "Industrial applications of model based predictive control," *Automatica*, vol. 29, no. 5, pp. 1251–1274, 1993.

[10] J. Maciejowski, *Predictive control: with constraints*. Prentice-Hall, Harlow, UK, 2002.

[11] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[12] C. Cutler and B. Ramaker, "Dynamic matrix control-a computer control algorithm," in *Proceedings of the Joint Automatic Control Conference*, 1980.

[13] J. Richalet, A. Rault, J. Testud, and J. Papon, "Model predictive heuristic control: Applications to industrial processes," *Automatica*, vol. 14, no. 5, pp. 413–428, 1978.

[14] D. Mayne, J. B. Rawlings, C. Rao, and P. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, June 2000.

[15] E. F. Camacho and C. A. Bordons, *Model Predictive Control in the Process Industry*. Secaucus, NJ, USA: Springer-Verlag, New York, 1997.

[16] D. W. Clarke, *Advances in Model-Based Predictive Control*. Oxford University Press, 1994.

[17] K. R. Muske and J. B. Rawlings, "Model predictive control with linear models," *AIChE Journal*, vol. 39, no. 2, pp. 262–287, 1993.

[18] S. Shin and S. Park, "GA-based predictive control for nonlinear processes," *Electronics Letters*, vol. 34, no. 20, pp. 1980–1981, Oct 1998.

[19] A. Alaniz, "Model predictive control with application to real-time hardware and a guided parafoil," Master's thesis, Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, Cambridge, MA, USA, 2004.

[20] X. Blasco, M. Martinez, J. Senent, and J. Sanchis, "Generalized predictive control using genetic algorithms (GAGPC): an application to control of a non-linear process with model uncertainty," in *Methodology and Tools in Knowledge-Based Systems*, ser. Lecture Notes in Computer Science. Springer, 1998, pp. 428–437.

[21] T. Kawabe and T. Tagami, "A real coded genetic algorithm for matrix inequality design approach of robust PID controller with two degrees of freedom," in *Proceedings of the 1997 IEEE International Symposium on Intelligent Control*, Jul 1997, pp. 119–124.

[22] R. Krohling, H. Jaschek, and J. Rey, "Designing PI/PID controllers for a motion control system based on genetic algorithms," in *Proceedings of the 1997 IEEE International Symposium on Intelligent Control*, Jul 1997, pp. 125–130.

[23] P. Angeline, "Using selection to improve particle swarm optimization," in *Proceedings of the 1998 IEEE International Conference on Evolutionary Computation, IEEE World Congress on Computational Intelligence.*, May 1998, pp. 84 –89.

[24] R. Krohling and J. Rey, "Design of optimal disturbance rejection PID controllers using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 1, pp. 78–82, Feb 2001.

[25] D. B. Fogel, *Evolutionary computation: toward a new philosophy of machine intelligence*. Piscataway, NJ, USA: IEEE Press, 1995.

[26] R. C. Eberhart and Y. Shi, "Comparison between genetic algorithms and particle swarm optimization," in *Proceedings of the 7th International Conference on Evolutionary Programming VII*, ser. EP '98. London, UK: Springer-Verlag, 1998, pp. 611–616.

[27] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, nov/dec 1995, pp. 1942–1948.

[28] Y. Shi and R. Eberhart, "A modified particle swarm optimizer," in *Proceedings on Evolutionary Computation, IEEE World Congress on Computational Intelligence.*, May 1998, pp. 69–73.

[29] H. Yoshida, K. Kawata, Y. Fukuyama, S. Takayama, and Y. Nakanishi, "A particle swarm optimization for reactive power and voltage control considering voltage security assessment," *IEEE Transactions on Power Systems*, vol. 15, no. 4, pp. 1232–1239, Nov 2000.

[30] D. W. Boeringer and D. H. Werner, "Particle swarm optimization versus genetic algorithms for phased array synthesis," *IEEE Transactions on Antennas and Propagation*, vol. 52, no. 3, pp. 771–779, March 2004.

[31] R. Hassan, B. Cohanim, O. de Weck, and G. Venter, "A comparison of particle swarm optimization and the genetic algorithm," in *46th AIAA, ASME, ASCE, AHS, ASC Structures, Structural Dynamics and Materials Conference*, Austin, USA, April 18-21 2005.

[32] M. R. AlRashidi and M. E. El-Hawary, "A survey of particle swarm optimization applications in electric power systems," *Evolutionary Computation, IEEE Transactions on*, vol. 13, no. 4, pp. 913 –918, aug. 2009.

[33] R. Eberhart and Y. Shi, "Particle swarm optimization: developments, applications and resources," in *Proceedings of the 2001 Congress on Evolutionary Computation*, vol. 1, 2001, pp. 81–86.

[34] X. Hu, Y. Shi, and R. Eberhart, "Recent advances in particle swarm," in *Congress on Evolutionary Computation, CEC2004*, vol. 1, June 2004, pp. 90–97.

[35] Y. del Valle, G. Venayagamoorthy, S. Mohagheghi, J.-C. Hernandez, and R. Harley, "Particle swarm optimization: Basic concepts, variants and applications in power systems," *IEEE Transactions on Evolutionary Computation*, vol. 12, no. 2, pp. 171–195, April 2008.

[36] M. Wachowiak, R. Smolikova, Y. Zheng, J. Zurada, and A. Elmaghraby, "An approach to multimodal biomedical image registration utilizing particle swarm optimization," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 289–301, June 2004.

[37] L. Messerschmidt and A. Engelbrecht, "Learning to play games using a PSO-based competitive learning approach," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 280–288, June 2004.

[38] N. Franken and A. Engelbrecht, "Particle swarm optimization approaches to coevolve strategies for the iterated prisoner's dilemma," *IEEE Transactions on Evolutionary Computation*, vol. 9, no. 6, pp. 562–579, Dec. 2005.

[39] X. Li and A. P. Engelbrecht, "Particle swarm optimization: an introduction and its recent developments," in *Proceedings of the 2007 GECCO conference companion on Genetic and evolutionary computation*, ser. GECCO '07. New York, USA: ACM, 2007, pp. 3391–3414.

[40] C. Coello, G. Pulido, and M. Lechuga, "Handling multiple objectives with particle swarm optimization," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 256–279, June 2004.

[41] Z.-L. Gaing, "Particle swarm optimization to solving the economic dispatch considering the generator constraints," *Power Systems, IEEE Transactions on*, vol. 18, no. 3, pp. 1187 – 1195, aug. 2003.

[42] B. Zhao, C. X. Guo, and Y. J. Cao, "Improved particle swam optimization algorithm for OPF problems," in *Power Systems Conference and Exposition, 2004. IEEE PES*, oct. 2004, pp. 233 – 238 vol.1.

[43] C.-M. Huang, C.-J. Huang, and M.-L. Wang, "A particle swarm optimization to identifying the armax model for short-term load forecasting," *Power Systems, IEEE Transactions on*, vol. 20, no. 2, pp. 1126 – 1133, may 2005.

[44] J.-B. Park, K.-S. Lee, J.-R. Shin, and K. Y. Lee, "A particle swarm optimization for economic dispatch with nonsmooth cost functions," *Power Systems, IEEE Transactions on*, vol. 20, no. 1, pp. 34 – 42, feb. 2005.

[45] W. Zhang and Y. Liu, "Reactive power optimization based on PSO in a practical power system," in *Power Engineering Society General Meeting, 2004. IEEE*, june 2004, pp. 239 – 243 Vol.1.

[46] M. Clerc and J. Kennedy, "The particle swarm - explosion, stability, and convergence in a multidimensional complex space," *Evolutionary Computation, IEEE Transactions on*, vol. 6, no. 1, pp. 58 –73, feb 2002.

[47] G. Coath and S. Halgamuge, "A comparison of constraint-handling methods for the application of particle swarm optimization to constrained nonlinear optimization problems," in *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, vol. 4, dec. 2003, pp. 2419 – 2425.

[48] A. I. El-Gallad, M. E. El-Hawary, and A. A. Sallam, "Swarming of intelligent particles for solving the nonlinear constrained optimization problem," *International Journal of Engineering Intelligent Systems*, vol. 9, no. 3, pp. 155–163, 2001.

[49] K. Yasuda, A. Ide, and N. Iwasaki, "Stability analysis of particle swarm optimization," in *Proceedings of The Fifth Metaheuristics International Conference*, 2003, pp. 341–346.

[50] R. Eberhart and J. Kennedy, "A new optimizer using particle swarm theory," in *Proceedings of the 6th IEEE International Symposium on Micro Machine and Human Science (MHS '95)*, Nagoya, Japan, October 1995, pp. 39–43.

[51] R. C. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence*, 1st ed., ser. The Morgan Kaufmann Series in Evolutionary Computation. San Francisco, USA: Morgan Kaufmann, 2001.

[52] A. P. Engelbrecht, *Computational Intelligence: An Introduction*, 2nd ed. Wiley Publishing, 2007.

[53] R. Eberhart, P. Simpson, and R. Dobbins, *Computational intelligence PC tools*. San Diego, CA, USA: Academic Press Professional, Inc., 1996.

[54] M. Omran, A. Salman, and A. Engelbrecht, "Image classification using particle swarm optimization," in *Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning*, 2002, pp. 370–374.

[55] Y. Shi and R. C. Eberhart, "Parameter selection in particle swarm optimization," in *Proceedings of the 7th International Conference on Evolutionary Programming VII*, ser. EP '98. London, UK: Springer-Verlag, 1998, pp. 591–600.

[56] H. Yoshida, Y. Fukuyama, S. Takayama, and Y. Nakanishi, "A particle swarm optimization for reactive power and voltage control in electric power systems considering voltage security assessment," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 6, 1999, pp. 497–502.

[57] P. Suganthan, "Particle swarm optimiser with neighbourhood operator," in *Proceedings of the Congress on Evolutionary Computation CEC 99.*, vol. 3, 1999, pp. 3 vol. (xxxvii+2348).

[58] A. Ratnaweera, S. Halgamuge, and H. Watson, "Particle swarm optimization with self-adaptive acceleration coefficients," in *Proceedings of the First International Conference on Fuzzy Systems and Knowledge Discovery*, 2003, pp. 264–268.

[59] S. Naka, T. Genji, T. Yura, and Y. Fukuyama, "Practical distribution state

estimation using hybrid particle swarm optimization," in *IEEE Power Engineering Society Winter Meeting*, vol. 2, 2001, pp. 815–820.

[60] J. J. E. Slotine and W. Li, *Applied Nonlinear Control*. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1991.

[61] The MathWorks, Inc. (2012, Dec) Simulink - simulation and model-based design. [Online]. Available: http://www.mathworks.com/products/simulink/

[62] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, USA: Cambridge University Press, 2004.

[63] A. P. Engelbrecht, "Scalability of a heterogeneous particle swarm optimizer," in *Swarm Intelligence (SIS), 2011 IEEE Symposium on*, april 2011, pp. 1 –8.

[64] A. Engelbrecht, "Heterogeneous particle swarm optimization," in *Swarm Intelligence*, ser. Lecture Notes in Computer Science, vol. 6234. Springer Berlin Heidelberg, 2010, pp. 191–202.

[65] B. J. Leonard and A. P. Engelbrecht, "Scalability study of particle swarm optimizers in dynamic environments," in *Swarm Intelligence*, ser. Lecture Notes in Computer Science, M. Dorigo, M. Birattari, C. Blum, A. L. Christensen, A. P. Engelbrecht, R. Gross, and T. Stützle, Eds., vol. 7461. Springer Berlin Heidelberg, 2012, pp. 121–132.

[66] S. Piccand, M. O'Neill, and J. Walker, "On the scalability of particle swarm optimisation," in *Evolutionary Computation, 2008. CEC 2008. (IEEE World Congress on Computational Intelligence). IEEE Congress on*, june 2008, pp. 2505 –2512.

[67] K. E. Parsopoulos and M. N. Vrahatis, "Particle swarm optimization method for constrained optimization problems," in *Intelligent Technologies – Theory and Applications: New Trends in Intelligent Technologies*. IOS Press, 2002, pp. 214–220.

[68] V. Tandon, H. El-Mounayri, and H. Kishawy, "NC end milling optimization using evolutionary computation," *International Journal of Machine Tools and Manufacture*, vol. 42, no. 5, pp. 595–605, 2002.

[69] F. Zhang and D. Xue, "An optimal concurrent design model using distributed product development life-cycle databases," in *Sixth International Conference on Computer Supported Cooperative Work in Design*, 2001, pp. 273–278.

[70] Y. Shi and R. Krohling, "Co-evolutionary particle swarm optimization to solve min-max problems," in *Proceedings of the 2002 Congress on Evolutionary Computation, CEC '02.*, vol. 2, 2002, pp. 1682–1687.

[71] E. Laskari, K. Parsopoulos, and M. Vrahatis, "Particle swarm optimization for integer programming," in *Proceedings of the Congress on Evolutionary Computation, CEC '02.*, vol. 2, 2002, pp. 1582–1587.

[72] X. Hu and R. Eberhart, "Solving constrained nonlinear optimization problems with particle swarm optimization," in *Proceedings of the Sixth World Multiconference on Systemics, Cybernetics and Informatics*, 2002.

[73] A. El-Gallad, M. El-Hawary, A. Sallam, and A. Kalas, "Enhancing the particle swarm optimizer via proper parameters selection," in *Canadian Conference on Electrical and Computer Engineering, IEEE CCECE*, vol. 2, 2002, pp. 792–797.

[74] G. Venter and J. Sobieszczanski-Sobieski, "Multidisciplinary optimization of a transport aircraft wing using particle swarm optimization," *Structural and Multidisciplinary Optimization*, vol. 26, pp. 121–131, 2004.

[75] G. Venter and J. Sobieszczanski-Sobieski, "Particle swarm optimization," *Journal for the American Institute of Aeronautics and Astronautics*, vol. 41, no. 8, pp. 1583–1589, 2003.

[76] J. C. Spall, "Stochastic optimization," in *Handbook of Computational Statistics*, J. Gentle, W. Härdle, and Y. Mori, Eds. Springer-Verlag, New York, 2004, ch. II.6, pp. 169–197.

[77] T. G. Kolda, R. M. Lewis, and V. Torczon, "Optimization by direct search: New perspectives on some classical and modern methods," *SIAM Review*, vol. 45, pp. 385–482, 2003.

[78] D. C. Karnopp, "Random search techniques for optimization problems," *Automatica*, vol. 1, pp. 111–121, 1963.

[79] B. Houska and H. J. Ferreau. (2012, Dec) ACADO toolkit: Automatic control and dynamic optimization. [Online]. Available: http://www.mathworks.com/products/simulink/

[80] B. Houska, H. J. Ferreau, and M. Diehl, "ACADO toolkit - an open-source framework for automatic control and dynamic optimization," *Optimal Control Applications and Methods*, vol. 32, no. 3, pp. 298–312, 2011.

[81] Paola Falugi, Eric Kerrigan and Eugene van Wyk. (2012, Dec) Imperial College London, optimal control software (ICLOCS). [Online]. Available: http://www.ee.ic.ac.uk/ICLOCS/

[82] A. Wächter, "An interior point algorithm for large-scale nonlinear optimization with applications in process engineering," Ph.D. dissertation, Carnegie Mellon University, 2002.

[83] A. Wächter and L. Biegler. (2012, Dec) IPOPT - an interior point optimizer. [Online]. Available: https://projects.coin-or.org/Ipopt

Towards Certifiable Autonomic Computing Systems Part I: A Consistent and Scalable System Design

Haffiz Shuaib and Richard John Anthony
Autonomics Research Group
School of Computing and Mathematical Sciences,The University of Greenwich.
Park Row, Greenwich, London SE10 9LS, UK
Email: haffiz.shuaib@yahoo.com, R.J.Anthony@gre.ac.uk

***Abstract*—Relative to currently deployed Information Technology (IT) systems, autonomic computing systems are expected to exhibit superior control/management behaviour and high adaptability, regardless of operational context. However, a means for measuring and certifying the self-management capabilities of these systems is lacking and as result, there is no way of assessing the trustworthiness of these systems. Two things are needed to begin to address the above. The first is a consistent structure for the autonomic computing system (ACS) and a consistent architecture for the autonomic computing manager (AM). The second is a set of metrics by which the operational characteristics of these systems are to be measured within the context of the targeted application domain.**

In this first part of a two-part paper, a biologically inspired architecture is proposed for the autonomic computing manager. The interfaces and messages by which this architecture communicates with objects within and those without are technically defined. Also discussed in this paper is the policy structure by which the autonomic manager is configured to sense contexts and effect changes in its managed environment. For the system framework, a tree structure together with its associated protocols is proposed, implemented and used as the basis for establishing administrative and security relationships between autonomic computing elements; for resolving management conflicts; for enforcing data integrity; for ensuring data availability and for providing mechanisms that aid system scalability, robustness and extensibility, while maintaining low system complexity. This framework is achieved using standards-based objects including the Lightweight Directory Access Protocol (LDAP), Policy Core Information Model (PCIM) and a significant number of Internet Engineering Task Force (IETF) Request For Comments (RFC) standard documents.

Keywords-Autonomic computing systems; Certification; Architecture; Intelligent Machine Design; LDAP;

I. Introduction

Information Technology (IT) systems are rapidly growing in complexity and are becoming more difficult to manage by the day. This growing complexity requires an increase in the number of expert human operators managing these systems. This in turn increases the cost associated with IT service management. Therefore, steadily replacing the human operator with machines that can carry out similar managerial functions is desirable. Apart from cost savings, this has the added benefit of allowing complex computing systems to evolve into even more complex ones with the associated value added service. The human operator will act as an overall guide to the system and should in no way constitute a technological bottleneck. However, such a computing platform must be verified and trusted before it is handed complex managerial duties. To accomplish this task, the internal components of the computing managers must be well understood, as well as their interactions with managed elements. The system of managers must be of low complexity, scalable, portable, secure and be able to efficiently and effectively accomplish the managerial tasks. Being able to assign a consistent measure of trust to these systems is also important. These are the challenges that need to be resolved by the autonomic computing research field.

Although this research field is about a decade old, the solutions to certifying autonomic computing systems (ACS), though urgently needed, have not been considered. Proposed solutions must tackle the certification problem from the twin angles of the architecture of the system and its component managers, as well as mechanisms for deriving quantitative and qualitative measures for the ACS. These solutions, when implemented and verified will lead to further acceptance of ACSs.

One of the difficulties associated with certification in this regard has to do with the inability to achieve a fair comparison between autonomic systems or elements from two or more vendors, as each may adopt a different system structure or element architecture. In order to address this difficulty, an architecture for autonomic systems and managers that enforces structure but is flexible is required. To that end, a three-layered architecture referred to as the Intelligent Machine Design (IMD) is co-opted and technically defined for autonomic computing managers (AMs). The general form of this architecture is based on observations of how humans or animals behave in terms of the way they perceive their immediate environment and effect changes as a result.

An autonomic computing system will typically consist of manager and managed elements. These elements must be able to co-exist and interact gracefully with one another within the system. However, versatile and standardized mechanisms that should aid proper management coordination within an autonomic computing system are nonexistent. Later in this paper, the requirements necessary for the above are identified, a system that relies solely on standardized protocols is proposed, and how this system meets each of the previously identified requirements is discussed.

This paper collates together the findings of a detailed research project whose roadmap can be found in [1] and more extensive details in [2].

The rest of this paper is organized as follows; In the next section, the state of the art as it relates to autonomic manager (AM) architecture is discussed. Also presented in this section, is an expression of the Intelligent Machine Design (IMD) architecture for AMs. Interfaces, event message types, valid configurations, policy object framework for the IMD are proposed and presented in Sections IV, V, VI and VII, respectively. A standard structure on which an ACS can be built upon is proposed and presented in Section IX. In Section VIII, the requirements for management coordination and efficient autonomic elements interactions in an ACS are set out. The solution to each of these requirements is presented in Section X. The conclusion follows in Section XI.

II. Autonomic Manager (AM) Architecture

An architectural standard is central to the process of the certification of an object. Any architecture that represents an autonomic

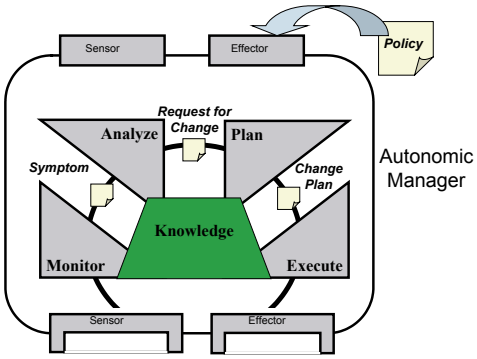


Fig. 1. IBM Autonomic MAPE Architecture [5]

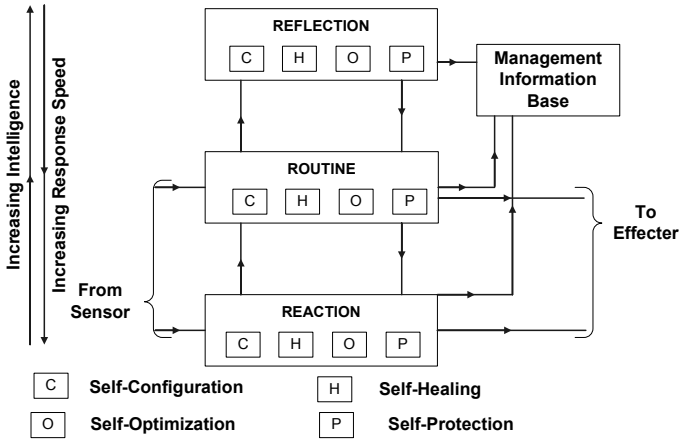


Fig. 2. An Autonomic Computing expression of the IMD

element e.g., an AM, must not be narrowly defined such that it precludes the ability for the element to evolve or cater for new use-cases. As noted in [3] and [4], the lack of an open standard is a challenge in the autonomic computing field. In this section, the most prevalent of all autonomic element architectures i.e., IBM's MAPE architecture is discussed. Drawbacks relating to this architecture are also discussed, and a certifiable alternative architecture with similar functionalities is presented.

A. Related Work

The IBM MAPE (Monitor, Analyze, Plan, Execute) architecture is a well known autonomic computing element architecture [5], and has been used as a reference for several autonomic computing systems. Systems that use the MAPE as a reference include; the web service host system proposed in [6], the self-adaptive service oriented system in [7] and the LOGO kit for data warehousing [8]. It has also been implemented in the Open Services Gateway initiative (OSGi) platform [9], applied to a Mobile Network Resource Management Architecture and several other projects [10][11].

The architecture consists of four main components which form a loop, as shown in Figure 1. The first of these components is the Monitor. Its main duty is to monitor the surrounding environment, including system resources. The output of this Monitor is used for making decisions at later stages of the loop. The second component i.e., the Analyze component, uses a number of algorithms to anticipate problems and possibly proffer solutions to these problems. The Planning component uses the information available to the autonomic system to choose which policies to execute. The Execution component, which is the fourth component, effects the most appropriate policy/policies chosen by the system. This executed policy may cause a change in the physical environment e.g., moving the arm of a robot, or simply pass instructions/information to another element, possibly an autonomic one. The input to the MAPE architecture comes from the sensory mechanism, while the effector mechanisms carry out the dictates of the machine.

While this architecture suffices for the purpose for which it was designed, it is ill-suited for certification purposes. This architecture has some limitations. For example, [12] considers it to be too narrowly defined to apply to some autonomic systems e.g., multi-agent systems. [13] points out that the loop in the MAPE architecture is vulnerable to failure, which in turn can precipitate the collapse of the management system all together. In addition to the above, there is no consensus as to whether the IBM MAPE architecture is a concrete architecture or a malleable concept. As a result, there are many different MAPE implementation permutations. A discussion of these divergent implementation views can be found in [1]. The current lack of a consistent architectural structure for the autonomic manager hampers the certification process.

Given that autonomic computing systems are biologically inspired, it follows that the manager architecture should also be similarly inspired, after all, AMs are supposed to steadily replace the human operator. This architecture must enforce structure without impeding innovation and it must allow for the separation of concerns i.e., the grouping of components with similar functionalities. This is the approach taken for the architecture presented in the next section.

B. Intelligent Machine Design (IMD)

The appeal of the Intelligent Machine Design (IMD) architecture [14] to autonomic computing systems is that it is closely related to the way intelligent biological systems work. The theory that underpins this architecture proceeds from the mechanisms by which animals and humans evaluate and effect changes in an environment, using their affect and cognitive abilities. Indeed, this architecture has been suggested as a generic framework on which, autonomic systems can be built upon [15]. While this architecture is mentioned in some autonomic computing literature, nothing concrete from a technical perspective has been achieved relative to IBM's architecture. Before describing this architecture in detail, it is necessary to specify what a Policy Rule is. A Policy Rule is the primary technical mechanism by which an AM effects changes in its managed environment given a specific context. A policy rule is made up of policy conditions, policy actions and other policy data that indicate how a policy condition is evaluated and how a policy action is to be executed. If the perceived state of the managed environment corresponds to the condition of a Policy rule, the AM executes the associated policy action accordingly. See Section VII for technical details on these Policy objects.

The IMD architecture is made up of three distinct layers i.e., Reaction (R1), Routine (R2) and the Reflection (R3) level (see Figure 2). Each layer is characterized by the following attributes; the amount of resources consumed, their ability to activate/inhibit the functionality of a connected layer and their ability to be activated or inhibited by another layer.

The lowest layer, the Reaction layer, is connected to the sensors (S) and effecters (E). When it receives a sensory stimulus, it responds relatively faster than the other two layers. The primary reason for this is that its internal mechanisms are simple, direct and hardwired

i.e., it has an automatic response to incoming signals. Technically, the Reaction layer implements a single policy rule for all received input signals. However, if the input signal is such that this solitary policy rule does not suffice, control is handed over to the Routine layer (R2). The Reaction layer can also be inhibited/activated by the Routine layer. It consumes the least amount of resources.

The Routine (mid-level) layer is more learned and skilled when compared to the Reaction layer. It is expected to have access to the working memory or the Management Information Base (MIB), which contain a number of policy rules that are executed based on context, knowledge and self-awareness. As a result, it is comparatively slower than the Reaction level. Its activities can be activated or inhibited by the Reflection layer. Its input comes from both the sensory mechanism and the Reflection layer. Its output goes to the effecter mechanism and the Reflection layer. When the Routine level is unable to find a suitable policy rule for an immediate objective, due to ambiguities between two or more existing policy rules or the lack of a policy rule thereof, it hands control over to the human administrator or the Reflection layer.

While the Routine level's primary objective is to deal with expected situations whether learned or hardwired, the Reflection level, which is the highest level, helps the machine deal with deviations from the norm. The Reflection level is able to deal with abnormal situations, using a combination of learning technologies (e.g., Artificial Neural Networks, genetic algorithms), partial reasoning algorithms (e.g., Fuzzy Logic, Bayesian reasoning), the machine's knowledge base, context and self-awareness. Technically, the Reflection Layer's ultimate aim, as it relates to autonomic computing systems, is to create and validate new policies at runtime that will be used at the Routine level. If the system is able to adapt to an unexpected situation as a result of the new policy rule, then the rule is stored in the MIB. This new rule can be called upon if the situation is encountered in the future. Thus, making a formerly abnormal situation a routine one. The process of 'reasoning' out a new policy rule makes the Reflection layer the largest consumer of computing resources. This also means it has the slowest response time of all three layers. The Routine layer is the input source and output destination for the Reflection layer. The Reflection layer can inhibit/activate the processes of the Routine layer through new policy rule definitions.

C. The IMD and The Four Cardinal Self-Management Properties

Notice in Figure 2, that all three layers of the IMD are able to action each of the four 'cardinal' autonomic management properties (self- configuration, healing, optimisation and protection). To demonstrate how this works, consider an optimization scenario where limited resources must be allocated between competing requests. In this scenario, when the number of requests go beyond a certain threshold then the system is in danger of collapsing e.g., a sudden build up of service requests, leading to service queue overflow and thus violation of Service Level Agreements (SLAs). Assume that the number of requests currently being handled by the system is near that threshold. On sensing that the threshold is about to be reached, an autonomic manager (AM) implementing the IMD as its architecture engages its Reaction layer. The self-optimization component of the Reaction layer immediately forces the AM to stop any further allocation of resources to requests. There is little or no intelligence involved in this action. The Reaction layer then informs the Routine layer of this action. The Routine layer needs to effect an action such that the requests with higher priorities (based on the organizational goals) are met. To do this, the Routine layer looks to its policy rule database or MIB, to find the most appropriate rule whose condition fits the context. If an appropriate policy rule that optimizes the use of the limited resource is found, its associated action is executed. The execution of this policy rule's action overrides the lock placed on the managed system by the Reaction layer. Note that this Routine layer adopts a more intelligent and fine-grained approach to solving the optimization problem.

It may be that the Routine layer is unable to find a suitable policy rule in the policy repository for this specific context. In a case like this, control is handed over to the Reflection layer. Keep in mind that the lock implemented by the Reaction layer is still in effect at this time. The Reflection layer 'deliberates' on the best combination of requests that can be granted access to the managed resources, while still ensuring that the system is stable and organizational goals are met given the current context. The Reflection layer will be expected to implement a utility function or an artificial intelligence algorithm for this optimization process. As soon as a solution is computed a new policy rule is created and added to the policy repository and the Routine layer is informed of same. The Routine layer is now at liberty to effect the new policy rule. Again, as soon as the action associated with the new rule is executed, the previous lock placed by the Reaction layer on system resource allocation is removed. The same principle applies to the other three self-management properties.

Note that the terms 'the machine' and 'IMD' are synonymous and are used interchangeably through out this paper.

III. An Autonomic Application Example

An autonomic application example is used to illustrate and put into context some of the technical details presented in this work. For this purpose, an application called Path Finder (PF) in which robots are guided by AMs to and fro between a base and a target on a gridded map. The objective of this application is to have robots accomplish as many round trips as is possible between the base and the target within a considered time. The robots can be moved once on a clock tick in one of four directions on the map i.e., Top, Bottom, Left or Right square. To carry out this task, the AM must deduce, through its sensory mechanism (S), how many of the four squares constitute a valid next move for the robot, given its current position. Using its artificial intelligence algorithm, the AM decides which of the valid or available squares is best for the robot's next move. When the decision is made, the effecter mechanism (E) moves the robot, accordingly. In autonomic parlance, the robot is the *Managed Resource/Element* while the AM is the *Manager Element*.

IV. Machine Interfaces

Four distinct types of interfaces are proposed for the IMD in this work. These interfaces are shown in Figure 3 and are labeled *I-1* - *I-4*. Each interface is discussed in terms of the kind and structure of information it allows through.

A. The I-1 Interface

The first interface, *I-1* connects the Reaction and Routine layers to the sensory input (S) of the machine. Within this work, information that comes through the sensory interface i.e., *I-1* is referred to as a '*Context*'. While it would be expected that different autonomic applications would implement different *Context*s, it is necessary to describe this input information in a standardized way. The reason for this is that as long as an AM complies with the standard, it will always be able to interpret a *Context*, irrespective of the target application. The IETF standard, RFC 2252 [16] is the means by which the structure of a *Context* is described. RFC 2252 provides a standard basis for which attributes of different data types are defined. Multiple

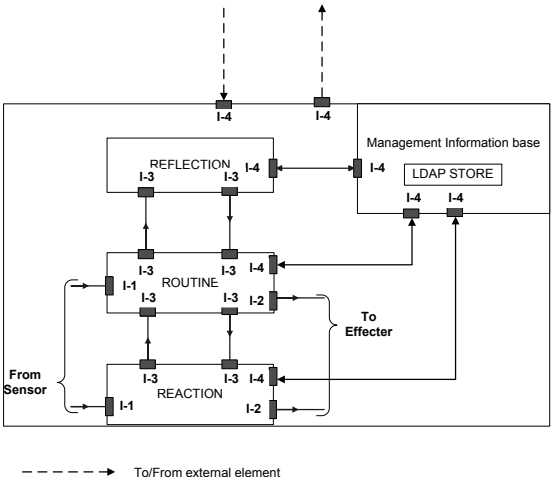


Fig. 3. Proposed interfaces for the IMD Architecture

```
objectclass ( 2.3.6.1.4.1.1.6863.6.1.909
NAME 'pfSensory' SUP Top STRUCTURAL
MUST (robotID $ topDirection $ bottomDirection $
rightDirection $ leftDirection) )
```

Fig. 4. RFC 2252 Compliant Object class for the PF *I-1*

attributes are grouped together under a structure called an object class. This RFC also mandates that all attributes and object classes are globally unique. This restriction ensures that no two applications will have the same *Context*. The object class in Figure 4 is an example of how the structure for the *Context* of the PF application described in Section III would look like.

In the figure, *2.3.6.1.4.1.1.6863.6.1.909* is a globally unique identifier for the object class called an Object ID (OID) and is assigned by the Internet Assigned Numbers Authority (IANA). Note that the OID in the figure above is fictitious. The name of the object class is *pfSensory* and it too must be globally unique. *SUP Top* means that the '*pfSensory*' class inherits the properties of another object class called *Top*. The *Top* object class is the parent class of all object classes defined in RFC 2252 and all sub-classes directly or indirectly inherit its properties. The *STRUCTURAL* Keyword simply means that *pfSensory* can be used as a stand-alone class. The *pfSensory* class consists of five mandatory attributes i.e., *robotID*, *topDirection*, *bottomDirection*, *rightDirection*, *leftDirection*. The *robotID* attribute is an integer value that contains the unique identifier of the robot. The other four attributes of this class i.e., *topDirection*, *bottomDirection*, *rightDirection* and *leftDirection* are Boolean values that are either True or False depending on whether a move to the Top, Bottom, Right or Left Square is valid, respectively. These attributes like the *pfSensory* object class must be defined in accordance with the rules set out in RFC 2252. The structure of these attributes will not be presented here. The interested reader should consult the RFC on how to go about this.

The *I-1* interface of AMs targeted at the *PF* application will only accept input information or *Context*s that are of the type *pfSensory*. With an information object class like this, multiple vendors can design AMs for this application without having to worry about compatibility issues. In addition, the contents of the *pfSensory* object class will have a globally consistent meaning, as it is compliant with the RFC 2252 standard.

```
objectclass ( 2.3.6.1.4.1.1.6863.6.1.910
NAME 'pfEffecter' SUP Top STRUCTURAL
MAY (topDirection $ bottomDirection
$ rightDirection $ leftDirection) )
```

Fig. 5. RFC 2252 Compliant Object class for the *I-2*

B. *The* I-2 *Interface*

Interface *I-2*, is used to instrument the physical environment or effect a change on the managed element. The instructions that lead to changes should be contained within the action code of the appropriate executing policy rule. The *I-2* interface on the other hand, sits between the AM and its effecter (E) mechanism, as shown in Figure 3. The information allowed on this interface must also be RFC 2252 compliant. An example of an object class for the I-2 interface for the *PF* is shown in Figure 5. This class consists of four optional boolean attributes viz; *topDirection*, *bottomDirection*, *rightDirection* and *leftDirection*. The ***MAY*** keyword in Figure 5 is what indicates that these attributes are optional. The AM creates an instance of this class and inserts the direction the robot is to be moved. This class instance is passed to the Effecter (E), which extracts the direction information and moves the robot in that direction accordingly. If more than one direction is specified in *pfEffecter* class instance, the effecter discards the information and the robot is not moved.

C. *The* I-3 *Interface*

The next interface, *I-3* is used for communication between layers of the IMD. Communication between layers is accomplished using a Machine Event Message (MEM). The exact technical details of the MEM are presented Section V. Recall from Section II-B, that a higher layer can modulate the response of a lower layer to an input stimulus or *Context*. This is accomplished through the *I-3* interface. Suffice it to say that this interface will only accept information that complies with the defined structure of the MEM.

D. *The* I-4 *Interface*

Before describing the *I-4* interface, it is instructive to briefly discuss the Lightweight Directory Access Protocol (LDAP) shown in Figure 3. In an autonomic computing system (ACS) and indeed any system, there is a need to have the ability to store and retrieve data information relating to management activities. Particularly, for ACSs, one needs to be able to store information relating to functional components within the autonomic domain e.g., managers, managed elements, policy objects, operational states of active elements, activity logs etc. All of the above will require a management information base (MIB). In this project, the Lightweight Directory Access Protocol (LDAP) defined in RFC 4510 [17] is the mechanism by which the MIB is realized. The LDAP is both a data storage/retrieval protocol as well as a communication protocol. As a data storage /retrieval protocol, it acts as a front-end to file storage systems that conform to the .X500 directory services. As a communication protocol, it runs atop the TCP/IP protocol stack. This provides an efficient, robust and secure link between any two autonomic elements.

The machine, therefore, uses its *I-4* interface to communicate with its LDAP compliant working memory or MIB and to communicate with external components, including a system-wide LDAP store, where available. As a result, only LDAP compliant operations are allowed on this interface. These operations are divided into three groups as delineated below;

1) **Interrogation operations**: Search and Compare.

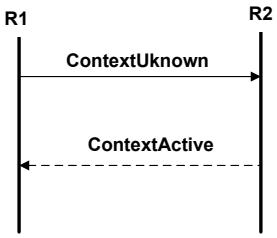


Fig. 6. Machine Event Message Exchange I

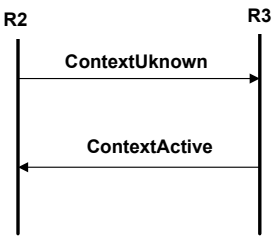


Fig. 7. Machine Event Message Exchange II

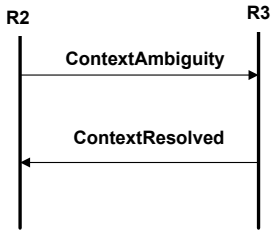


Fig. 8. Machine Event Message Exchange III

2) **Update operations**: Add, Delete, Modify and Rename.
3) **Authentication and control operations**: Bind, Unbind and Abandon.

Authentication and control operations are used to set up and tear down administrative and security relationships between autonomic elements. The retrieval of relevant information is based on the functionalities exposed by the *Interrogation operations*. Finally, the *Update operations* is used to carry out functions for which they are appropriately named.

V. Machine Event Messages

Messages exchanged between layers are called machine event messages (MEM) and are four-tuple objects that take the general form;

$$MEM = \begin{cases} eventType \\ eventID \\ policyRuleDNList \\ Context \end{cases}$$

The first component, *eventType*, of a MEM identifies what type of message a layer is signalling. Four types of message events are identified for this machine and they are;

$$eventType = \begin{cases} contextUknown \\ contextAmbiguity \\ contextActive \\ contextResolved \end{cases}$$

The purpose of each of these four event types are described later.

The *eventID* is a unique integer identifier for the MEM and the '*policyRuleDNList*' is a list that contains the identities or Distinguished Names (*DNs*) of policy rules. *Context* is the information retrieved from the *I-1* interface (see Section IV-A). The valid message exchange process between layers of the machine is described using Figures 6, 7 and 8.

Recall from Section II-B that (1) when the Reaction Layer (R1) is unable to deal with an incoming *Context* or signal, it hands control over to the Routine layer (R2) and (2) that R2 can regulate R1's response to incoming *Context*. The process is accomplished using the message sequence chart shown in Figure 6. If the R1 layer can handle an incoming *Context*, it simply executes the action contained within its singular policy rule, otherwise, it creates an MEM. The event message will have a unique integer value inserted into the *eventID* field. The *eventType* of the MEM is set to *contextUnknown*. The incoming *Context* is inserted into the *Context* field of this MEM. The *policyRuleDNList* is left empty. R1 hands control over to R2 by sending this newly created MEM to the R2 layer through the connecting *I-3* interface (see Section IV-C). On receipt of the MEM, the R2 layer retrieves the *Context* and uses this to search for the most suitable policy rule from its policy repository i.e., LDAP store. If found, the dictates of the action associated with the policy rule is passed to the effecter for execution. Note that the R2 layer will reject a MEM from R1 that contains an eventType other than *contextUnknown*. In order to modulate the response of the R1 layer to incoming signals, the R2 layer replaces the policy rule currently active in the R1 layer. To do this, R2 creates an MEM with the *evenType* set to *contextActive* and the name of the new active policy rule is inserted into the *policyRuleDNList*. As soon as the R1 layer receives this event message, it replaces its current active rule with that contained in the received *policyRuleDNList*.

On receiving a *Context* directly from the Sensory input (S) through the *I-1* interface or from a MEM created by R1, R2 may be unable to find a policy rule in the repository that matches the received *Context*. As discussed in Section II-B, in order to resolve this problem R2 must engage R3. To this end, R2 creates a MEM with its *eventType* set to *contextUknown* and the context field set to the received *Context*. Let this newly created MEM be called MEM1. This event message is transmitted to R3 through the I-3 interface. R3, using its implemented artificial intelligence algorithm, will attempt to create a new policy rule for this unknown *Context*. If successful, the new policy rule is written to the repository through R3's *I-4* interface. A new MEM with *eventType* set to *contextActive* is created and the distinguished name (DN) of the new policy rule is added to the *policyRuleDNList*. Let this MEM be called MEM2. Since MEM2 is a response to MEM1, both will share a similar *eventID* field value. When R2 receives MEM2 , it checks the *eventID* field to make sure it is a response to a previously sent MEM. If it is not, MEM2 is discarded and no action is taken for that *Context*. If the *eventID* field is a match to the *eventID* of a previously sent MEM, the policy rule in the *policyRuleDNList* is extracted and its associated action executed by the effecter (E) for that *Context*. If this *Context* is encountered in the future, there will be no need for R2 to reengage R3, as a matching policy rule has already been created previously. The message sequence for the above is depicted in Figure 7.

Consider a scenario where R2 finds that two or more policy rules in the repository match a particular *Context*. This uncertainty must be resolved by R3 (see Section II-B), this interaction is illustrated in Figure 8. In an instance like this, R2 creates MEM3 with

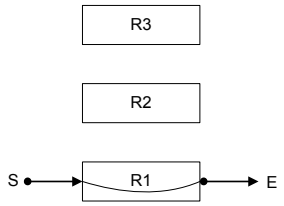


Fig. 9. Configuration I ($R1 \leftrightarrows \emptyset \leftrightarrows \emptyset$)

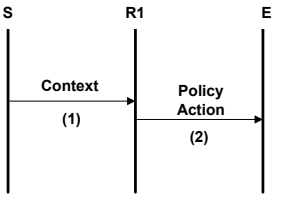


Fig. 10. Message Sequence Chart (Config. I)

eventType set to *ContextAmbiguity* and inserts the *Context* into the Context field of MME3. The DN of all rules that match the *Context* are added to the *policyRuleDNList* of MEM3. When R3 receives MEM3, it attempts to resolve the policy rule conflict. If it does, it generates MEM4, sets the *eventType* to *ContextResolved* and inserts the *eventID* of MEM3 into that of MEM4. The preferred policy rule in MEM3's *policyRuleDNList* is added to the *policyRuleDNList* of MEM4. MEM4 is sent to R2. On receipt, R2 will attempt to match the eventID of MEM4 to that of MEM3. If they are not a match, MEM4 is discarded. Otherwise, R2 executes the action contained within the policy rule in MEM4's *policyRuleDNList*. The interactions shown in Figures 7 and 8 are the means by which R3 regulates the activities of R2.

VI. Machine Configurations

There may be instances where a targeted autonomic application domain can do without the functionality of any one of the three layers of the IMD, for example, the R1 layer may or may not be needed. In another application example, the R3 layer may not be needed, if the targeted application does not require an artificial intelligence algorithm to determine behaviour for new or unexpected situations. As a result, the structure of the IMD lends itself to several layer configurations, five to be precise, depending on the autonomic application. These five configurations are governed by two rules.

1) **Rule 1**: A configuration must have at least an R1 or R2 layer. Observe from Figure 2 that only the Reaction (R1) and Routine (R2) layers have access to the sensor and effecter mechanisms. These are the means by which an IMD-compliant AM perceives and effects changes on the managed system. Without at least one of these two layers, the AM is ineffective.

2) **Rule 2**: This rule has to do with the presence of the R3 layer. If the R3 layer is present, then the R2 layer must also be present. Recall from Sections II-B and V, that the R3 layer resolves conflicts if two or more policy rules match a particular *Context*. Only R2 is able to detect rule conflicts, as R1 only implements a single policy rule.

Based on these two rules, the five valid machine configurations of the IMD and their allowed event message sequences are presented in Sections VI-A-VI-E.

A. Machine Configuration I

In the first valid configuration (shown in Figure 9), R2 and R3 are not in commission, meaning that R1 is the only active layer, giving rise to the $R1 \leftrightarrows \emptyset \leftrightarrows \emptyset$ configuration. Where $\leftrightarrows$ represents the bidirectional *I-3* interfaces that connect the layers (see Figure 3). R1 receives input from the sensory object (S). This input forms the current *Context*. If the conditions associated with the singular policy rule in R1 match this *Context*, R1 passes the associated policy actions to the effecter object (E) for execution. If not, control is passed to the human operator of the application.

The possible message sequence for the machine when the human operator is not involved is shown in Figure 10and represented by Expression (1).

$$(1) \mapsto (2) \tag{1}$$

The symbol $\mapsto$ in Expression (1) indicates the execution sequence from the sensing of a *Context* to the effecting of a change in the managed environment.

B. Machine Configuration II

Configuration II (Figure 11) i.e., $R1 \leftrightarrows R2 \leftrightarrows \emptyset$, assumes that the R3 layer is not needed for the targeted application domain. There are two possible message sequences for this configuration. These are shown in Expressions (2) and (3) and depicted in Figure 12.

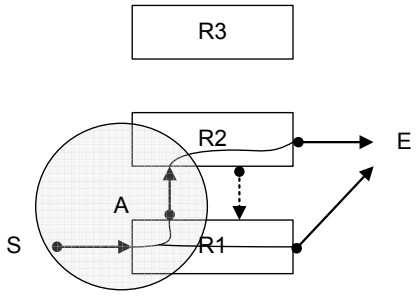


Fig. 11. Configuration II ($R1 \leftrightarrows R2 \leftrightarrows \emptyset$)

$$(1) \mapsto (2) \tag{2}$$

$$(1) \mapsto (3) \mapsto (4) \tag{3}$$

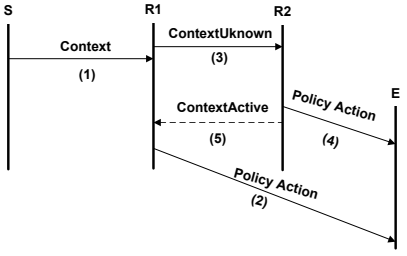


Fig. 12. Message Sequence Chart (Config. II)

Notice that Expression (2) corresponds to Expression (1) in Section VI-A. This indicates that Configuration II is an extension of Configuration I. If possible, the portion circled and labeled '**A**' in Figure 11 should be implemented as a single self-contained function. As is shown later, this function is reusable when a machine with this configuration needs to be extended. The dotted arrows shown in

Figures 11 and 12 represent instances where R2 changes the policy rule implemented in R1.

C. Machine Configuration III

All three layers of the machine in Configuration III are active as shown in Figure 13. This configuration allows for three different message sequence depending on the *Context* (see Expressions (4), (5), (6) and Figure 14).

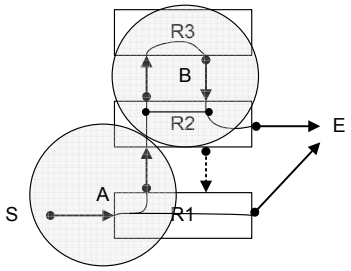


Fig. 13. Configuration III ($R1 \leftrightarrows R2 \leftrightarrows R3$)

$$(1) \mapsto (2) \tag{4}$$

$$(1) \mapsto (3) \mapsto (4) \tag{5}$$

$$(1) \mapsto (3) \mapsto (6) \mapsto (7) \mapsto (4) \tag{6}$$

Observe that the message sequence represented by Expressions (4) and (5) are also present in the Expressions of Configuration II. Again, this is a concrete expression of the fact that Configuration III simply an extension of Configuration II.

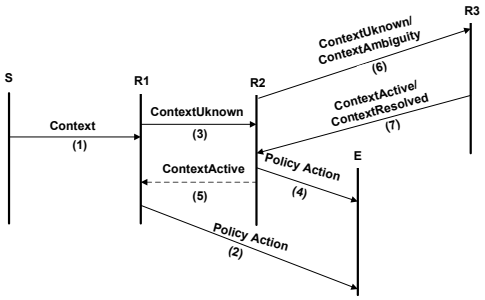


Fig. 14. Message Sequence Chart (Config. III)

Still on the theme of extensibility, notice that Figure 13 also has a portion circled '**A**', as did Figure 11 of Configuration II. If for any reason a machine with Configuration II needs to be extended to Configuration III, the self-contained function that implements the circled portion '**A**' of Figures 11 and 13 need not be rewritten, as it can be used as is. In a similar vain, the portion circled and labeled '**B**' in Figure 13 should also be implemented in a single self-contained function, as it can be reused when a need arises to transition from Configuration III to Configuration V.

D. Machine Configuration IV

In configuration IV, only the R2 layer is active as shown in Figure 15. This means that as soon as a *Context* is sensed, the matching policy rule is found and its associated policy action is passed to the Effecter (E). The only message sequence allowed for this configuration is straight forward as laid out in Figure 16 and Expression (7).

$$(3) \mapsto (4) \tag{7}$$

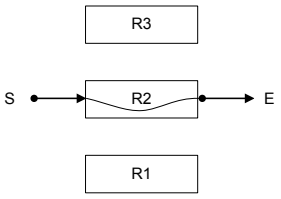


Fig. 15. Configuration IV ($\emptyset \leftrightarrows R2 \leftrightarrows \emptyset$)

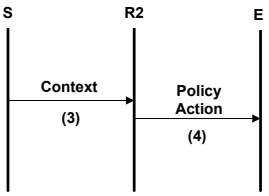


Fig. 16. Message Sequence Chart (Config. IV)

E. Machine Configuration V

From Figures 17 and 18 and Expressions (8) and (9), it is clear that Configuration IV is a subset of Configuration V.

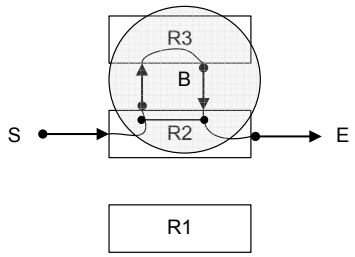


Fig. 17. Configuration V ($\emptyset \leftrightarrows R2 \leftrightarrows R3$)

$$(3) \mapsto (4) \tag{8}$$

$$(3) \mapsto (6) \mapsto (7) \mapsto (4) \tag{9}$$

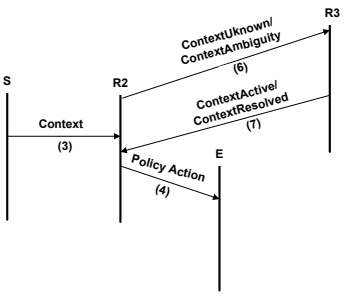


Fig. 18. Message Sequence Chart (Config. V)

This configuration is for application instances where the R1 layer may not be needed. Therefore, like Configuration IV, there is no

need to attempt to change the policy rule implemented in R1. The circled portion labeled '**B**' in Figure 13 of Configuration III is also present in Figure 17. As discussed in Section VI-C, implementing the portion labeled '**B**' as a self-contained function eases the process of extensibility, if required at some future time.

VII. Machine Policy Object Framework

In Section II-B, policy rules, conditions and actions for the AM were introduced but without their proper technical structures. In this section, these policy objects are discussed from a technical perspective. The Policy Core Information Model (PCIM) in RFC 3060 [18] and its update defined in RFC 3460 [19] is the primary framework by which an IMD- compliant AM and by extension an ACS are able to process a received *Context*, select the best action as a result and aid some system management functions.

The main appeal of the PCIM framework is that it is well established, in that it is standardized and used in a number of computing management related fields, for example the management of computer networks. Apart from the fact that this framework outlines the structures by which policy objects are defined, it also enforces type safety, which allows for ease of parsing and automation. And this it does without restricting problem-specific applicability. The PCIM framework is able to achieve all of the above by defining the processes and associated schemas by which already defined policy object classes are to be encapsulated, extended or reused. The PCIM defines a number of policy objects but only four of these i.e., policy rules, policy conditions, policy actions and policy role collection are relevant to this work. Each of these policy objects implements an RFC 2252 compliant object class and consists of a number of similarly compliant object attributes. The object classes together with their associated attributes govern how specific policy objects are interpreted when read and/or executed. The classes and attributes of the four relevant policy objects are discussed further in the subsections that follow.

Note that this section is not meant to be an exhaustive discussion of the PCIM framework and its dependencies, as this information spans more than 10 Internet Engineering Task Force (IETF) RFCs or standards. However, it is detailed enough to support the core ideas discussed, leaving out extraneous information. In addition, when a policy object is introduced, the containing RFC is mentioned along side.

A. Policy Rule

A policy rule object is the means by which a condition or set of conditions is/are associated with an action or set of actions. According to RFC 3060/3460, it is not necessary for a policy rule to have an associated condition or action. However, in this work, all policy rules have conditions and corresponding actions. A policy rule is realized through the object class called *PolicyRule*. This class consists of 10 optional attributes. Only seven of these attributes are relevant to this work and they are discussed here. The first two attributes of the *PolicyRule* class to be dealt with are the *ConditionList* and the *ActionList*. The *ConditionList* contains the unique identities of the conditions that are to be evaluated when the policy rule is invoked. The *ActionList*, likewise, contains the unique identifiers of actions that would be executed if all the associated conditions evaluate to true. An attribute called the *ConditionListType* determines how the conditions of a policy rule are to be evaluated. This attribute specifies two types of condition evaluation procedures, namely; Disjunctive Normal Form (DNF) and Conjunctive Normal Form (CNF). In order to describe how the *DNF* and *CNF* apply to condition evaluation, it is necessary to state here that one or more conditions can be assembled under a single group number and a policy rule may be associated with more than one group of conditions. Assume that an instance of the policy rule class exists, such that it is made up of three groups of conditions. The first, i.e., Group 1 contains conditions C1 and C2, Group 2 contains C3 and C4 and Group 3 contains C5 and C6. If the *ConditionListType* is set to *DNF* (which is the default), the conditions are evaluated thus;

(C1 AND C2) OR (C3 AND C4) OR (C5 AND C6)

If *ConditionListType* = *CNF*, then

(C1 OR C2) AND (C3 OR C4) AND (C5 OR C6)

The *PolicyRule* class has an attribute called *Enabled*. This attribute can take one of three values i.e., *enabled*, *disabled* and *enabledForDebug*. If this attribute is set to *enabled* and if the associated conditions evaluate to true, the actions are executed. If it is set to *disabled*, the conditions are not evaluated and actions not executed. Lastly, if it is set to *enabledForDebug*, the conditions are evaluated but the actions are not executed.

pcimRuleSequencedActions attribute contains a list of integers that indicate the relative execution order of the policy actions associated with a *PolicyRule*. The values in this list are obtained from the *ActionOrder* attribute of the associated policy actions (see Section VII-C). The *Mandatory* attribute of the *PolicyRule* object class specifies the order, in which the policy actions associated with a policy rule are to be executed or interpreted. The allowed values for the *Mandatory* attribute are *mandatory*, *recommended* and *dontCare*. If *Mandatory* = *mandatory*, then the action order must be enforced, otherwise none of the actions should be executed. if *Mandatory* = *recommended*, the machine will attempt to execute the actions based on their order. If this fails, any other order may be attempted. If *Mandatory* = *dontCare*, the actions are executed in any order on the first try. In PCIM, managed elements can be grouped under a single named role. The name of this role can be added to a policy rule's *policyRoles* attribute. Every time this policy rule executes an action, the action impacts all managed elements pointed to by the content of its *policyRoles* attribute. The last attribute of the *PolicyRule* class discussed here is the *PolicyRuleName*. This attribute should ordinarily uniquely identify an instance of a policy rule object.

To store a defined policy rule instance in an LDAP store or MIB (see Section IV-D), the rule instance must follow the schema structure *pcelsRuleInstance* defined in RFCs 4104 [20] and 3703 [21]. These RFCs list this schema's globally unique identifier.

B. Policy Condition

A policy condition is defined by its object class called *PolicyCondition*. This is an abstract class that cannot be instantiated directly. It consists of four sub-classes, i.e., *PolicyTimePeriodCondition* , *SimplePolicyCondition*, *CompoundPolicyCondition* and *VendorPolicyCondition*. All of these sub-classes, save the last one are standardized. The *VendorPolicyCondition* is of most relevance to this work. This sub-class was created to allow for the definition of domain specific conditions that can be associated with policy rules. In other words, an instance of the ***PolicyCondition*** object can be applied to a vendor specific device through the *VendorPolicyCondition* sub-class.

According to RFC 4104 and 3703, creating a vendor specific condition and associating it with a policy rule is a four-step process.

1) First, the vendor must define the structure and interpretation of the input signal from the device. The defined structure of the signal is based on RFC 2252.
2) A schema for the vendor specific policy condition and its associated attributes must be defined according RFC 2252.
3) The defined vendor specific condition must then be coupled with an instance of what is called a policy rule association class.
4) The unique identifier of the instance of the association class is added to the policy rule's *ConditionList* attribute described in Section VII-A.

The PF application discussed in Section III is used to illustrate this four-step process. Step 1 has already been dealt with for the PF application in Section IV-A.

```
objectclass ( 2.3.6.1.4.1.1.6863.6.1.911
NAME 'pfCondition'
SUP pcimConditionVendorAuxClass AUXILIARY
MUST (isActive $ isValidMove $ topDirection
$ bottomDirection $ rightDirection $ leftDirection) )
```

Fig. 19. RFC 2252 Compliant PF Vendor Specific condition class

The second step in this process relates to defining the actual vendor policy condition. The schema for the vendor specific condition for the PF application is shown in Figure 19. From the figure it can be seen that the *pfCondition* class is derived from the *pcimCondition-VendorAuxClass* class, which is defined in RFC 3703. This vendor class has 7 attributes. The *isActive* attribute checks that the robot has been instantiated and is currently active. This attribute is always set to True. Attribute *isValidMove* verifies that the robot has not been moved on the current clock tick. Recall that a robot is moved at most once at the tick of the clock. The *isValidMove* attribute is also always set to True. The *topDirection*, *bottomDirection*, *rightDirection* and *leftDirection* attributes are similar to those discussed in Section IV-A. The *AUXILIARY* keyword indicates that the *pfCondition* class is not a stand-alone class and must be coupled with another class, which must be *STRUCTURAL* (see Section IV-A). In other words, its identity is drawn from the Structural class. This is the basis for Step 3 discussed later.

Assume that an AM maintains two variables for each robot i.e., *rActive* and *rMoved*. The *rActive* variable indicates the active state of the robot and the variable *rMoved* is either True or False depending on whether the robot has been moved in the current clock tick. Let *iSensory* be an instance of the *pfSensory* class (see Section IV-A) containing current information regarding a robot and its valid positions for the next move. An example condition evaluation code within an AM is shown in Figure 20.

```
isActive == rActive AND isValidMove == rMoved
AND (topDirection == iSensory.topDirection OR
bottomDirection == iSensory.bottomDirection OR
rightDirection == iSensory.rightDirection OR
leftDirection == iSensory. leftDirection)
```

Fig. 20. Condition evaluation code example

If the condition evaluates to True then all policy rules with matching conditions are equally applicable to the extant *Context*.

In the third step, the AUXILIARY *pfCondition* class must be coupled with a *STRUCTURAL* class called *pcelsConditionAssociation* class (as explained above). *pcelsConditionAssociation* is defined in RFC 4104 and its schema is shown in Figure 21.

```
objectclass ( 1.3.6.1.1.9.1.9
NAME 'pcelsConditionAssociation'
SUP pcimRuleConditionAssociation STRUCTURAL
MUST ( pcimConditionGroupNumber
$ pcimConditionNegated )
MAY ( pcimConditionName $ pcimConditionDN ) )
```

Fig. 21. RFC 2252 Compliant pcelsConditionAssociation class

The attribute *pcimConditionGroupNumber* is the group number to which the condition belongs. The *pcimConditionNegated* attribute indicates whether the condition should be negated before it is evaluated. A condition may have a condition name assigned to the attribute *pcimConditionName*, hence the use of the MAY keyword. Apart from the condition name, a DN or Distinguished Name may also identify the defined condition. The DN is assigned to the attribute *pcimConditionDN*. Coupling an *AUXILIARY* class to a *STRUCTURAL* class is necessary because an *AUXILIARY* class cannot be instantiated directly. When a *STRUCTURAL* class is instantiated, the attached *AUXILIARY* class is also instantiated but indirectly. By coupling the defined *pfCondition* to the *pcelsConditionAssociation* class, the attributes of the former are included with the attributes of the latter. A machine reading this condition instance sees only the *pcelsConditionAssociation* class and not the *pfCondition*. However, due to coupling the machine is able to read the attributes of the *pfCondition* object class.

In the fourth and final step, the DN of the instance of the coupled *pcelsConditionAssociation* is added to the *ConditionList* attribute of an instance of a *PolicyRule* class.

C. Policy Action

The creation of a vendor-specific policy action object class follows the same four-step process used to create and associate a policy condition to a policy rule. The first step in this case, is to create an object class by which information going to the effecter through the *I-2* interface must be an instance of. This was done for the PF application in Section IV-B. In the next step, the PF vendor policy action *AUXILIARY* object class has to be specified. An example is shown in Figure 22.

```
objectclass ( 2.3.6.1.4.1.1.6863.6.1.912
NAME 'pfAction'
SUP pcimActionAuxClass AUXILIARY
MUST functionID )
```

Fig. 22. 2252 Compliant PF Vendor Specific action class

It consists of a compulsory solitary attribute called *functionID* and the class is derived from the *pcimActionAuxClass* defined in RFC 3703. This *functionID* attribute is of data type string and it points to the function that creates an instance of the *pfEffecter* object class defined in Section IV-B. Once created, the instance of the *pfEffecter* class is passed to the effecter (E), which then moves the robot in the indicated direction.

For the third step, the *AUXILIARY pfAction* class is attached to the *STRUCTURAL pcelsActionAssociation* defined in RFC 4104. The compulsory *pcimActionOrder* attribute shown in Figure 23 indicates the relative execution order of a policy action, in a policy rule

```
objectclass ( 1.3.6.1.1.6.1.10
NAME 'pcelsActionAssociation'
SUP pcimRuleActionAssociation STRUCTURAL
MUST ( pcimActionOrder )
MAY ( pcimActionName $ pcimActionDN ) )
```

Fig. 23. RFC 2252 Compliant pcelsActionAssociation class

that consists of more than one policy action. The optional attributes *pcimActionName* and *pcimActionDN* hold the name of the action and the action's Distinguished Name (DN), respectively.

In the final step, the DN of the coupled instance of the *pcelsActionAssociation* is added to the *ActionList* attribute of an instance of the policy rule object. When all the conditions associated with a policy rule evaluate to true, the associated policy action is retrieved and the content of its *functionID* attribute is passed to the effecter for execution.

D. Policy Role Collection

Unlike the other three policy objects discussed previously, the Policy Role Collection object is simply an administrative unit that groups a number of related managed elements on which a single rule is applicable. Note that a named Policy Role Collection instance is associated to a specific policy rule using the *policyRoles* attribute of that rule (see Section VII-A). In other words, the *policyRoles* attribute contains the name of an instance of a Role Collection. If the conditions associated with a policy rule evaluate to true, then the associated policy actions are applied to all managed elements pointed to by the *policyRoles* attribute of the rule. The Policy Role Collection object is schematically represented by the object class *pcelsRoleCollection* defined in RFC 4104.

VIII. Requirements for Management Coordination in ACSs

As a testament to the need for the ability to coordinate and manage autonomic element interactions, autonomic computing literature is replete with instances or scenarios where several autonomic elements must interact to achieve a common goal. In [22], the autonomic managers communicate indirectly with one another using the system variables repository. If a manager were to fail, other managers reading this repository take over the responsibilities of the failed one. Other research works take a more direct approach to autonomic manager interaction. In [23] and [24], the communication between managers is peer-to-peer, while [25], [26], [27], [28] and [29] adopt a hierarchical system for manager interactions. These works either lack a formal definition of the mechanisms by which these autonomic managers interact, or where defined, these mechanisms were highly specific to the system in question, thus preventing wide applicability and reusability.

Notwithstanding the lack of a formal framework that addresses issues relating to autonomic element interoperability, attempts have been made to specify certain requirements that should be met if interoperability is to be made possible. For example, [30] argues that the mechanisms that define interoperability between autonomic elements must be reusable to limit complexities i.e., it must be generic enough to capture all communications across the board. [3] mentions the need for a name service registry for autonomic elements, a system interaction broker and a negotiator as necessary components for autonomic element interaction. Also required is a need for standardized communication interfaces between autonomic elements to ensure interactions are well documented and secure [31], [12]. Based on some of the information contained in these works, the following eight requirements are proposed for effective management coordination and element interaction in ACSs;

1) **Administrative relationships**: A means to establish proper administrative relationships should exist. This way the sphere of influence of autonomic managers is clearly defined. This requirement is necessary to solve problems associated with operational conflicts. Also included within this requirement, is the need to define clear procedures for security relationships between elements in an ACS.
2) **Conflict Resolution Mechanism**: A conflict resolution mechanism must exist if two or more managers are able to simultaneously effect changes on the same resource.
3) **Monitoring Autonomic Elements**: A means must exist to query the internal state of an autonomic element. This is taken for granted when an AM might inquire as to the current state of an ME (e.g., start, stop and resume). Nevertheless, it may be necessary for an AM determine whether another AM is in a suitable operational state to allow for element interaction.
4) **Grant and Request Services**: For Requirement 3 to be possible, a mechanism for requesting and granting services must exist. For instance, an AM might need to understand the context in which a peer AM took an action. Requesting contextual information is within the remit of this requirement.
5) **Remote Policy Object Communication**: Following from Requirement 5, queries and associated responses must be transparent, regardless of the relative physical location of the AMs and MEs. In this case, an appropriate standardized communication protocol must exist to satisfy this requirement.
6) **Policy Object sharing**: If two or more AMs implement the same policy rule or if two or more MEs are instructed using the same policies, then an administrative mechanism (e.g., a well defined policy repository) for policy sharing should exist.
7) **A Policy Rule Selection Mechanism**: A structure to support the selection of the best policy for a given context should be available to a multi-policy system.
8) **Low complexity and Reusability**: Finally, the framework must be reusable across a broad spectrum of autonomic application domains without increasing its complexity.

The ways in which these requirements are met by the technical proposals in this work are presented in Section X.

IX. A System Architecture for ACSs

It is necessary to describe how the structure and build of an autonomic computing system (ACS) is approached in this work. The mechanisms discussed herein contribute to the solutions for the management coordination requirements set out in the last section.

In addition to the LDAP being the basis for an MIB (see Section IV-D), it is also the structure on which autonomic computing elements within an administrative domain are brought together to form an ACS. The mechanisms of the LDAP that provide for the core structure of the system, that ensure data integrity, security and availability are discussed here.

It may be noteworthy to state here that as the LDAP is an IETF standard, there are several implementations available. However, for this project, the *openLDAP* platform is the preferred choice. The reason for this is that in addition to being able to run on multiple operating systems, it is free, open source and implements a non-propriety license.

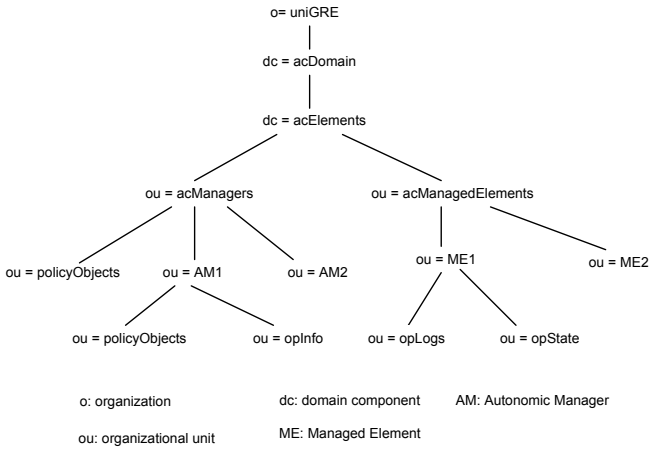


Fig. 24. Example DIT for an Autonomic Computing System

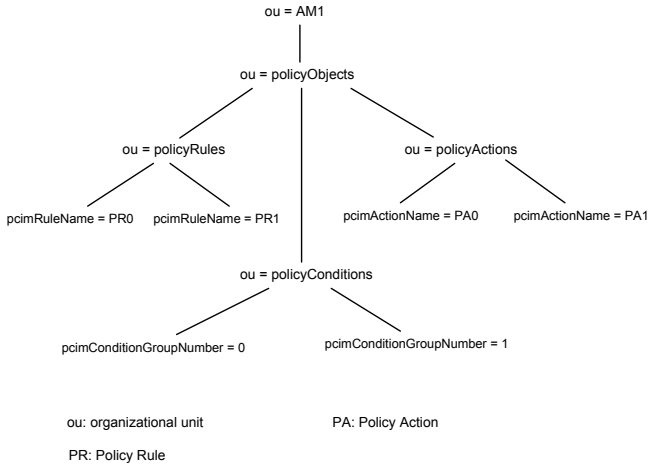


Fig. 25. AM1 Policy Object Branch

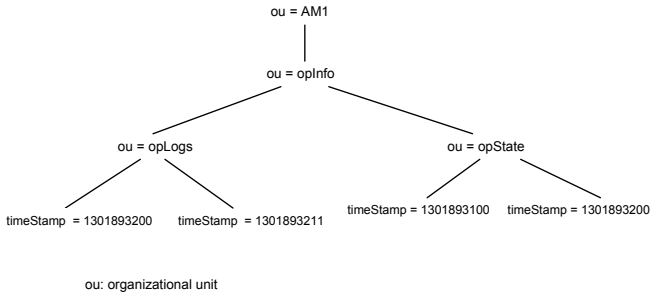


Fig. 26. AM1 Operational Information (opInfo) Branch

A. Core Structure

The LDAP stores data in a structure referred to as the Directory Information Tree (DIT). An example DIT implemented in this project is shown in Figure 24, with branches in Figures 25 and 26. At the root of the tree is *o=uniGre*. The *uniGre* (or University of Greenwich) component of this root entry is the name of the organization that owns the directory. The *o* component of this root entry is the name of the object class that defines the rules governing the naming of organizations in a DIT. This object class is defined in RFC 2256 [32]. Put more succinctly, the entry 'o=uniGre' means that the name of the organization is *uniGre* and this name conforms to the *o* object class.

At the next level of the DIT is the name of the domain within the organization i.e., *acDomain*, which stands for autonomic computing domain. The domain name conforms to the domain component (*dc*) object class defined in RFC 2247 [33]. The *acDomain* has a single branch called *acElements*. The *acElements* domain is split between autonomic computing managers and managed elements. These two branches conform to the organizational unit (*ou*) object class defined in RFC 2256. The *ou* object is a container that holds a number of other object classes. All managed devices in the organization are placed under the *acManagedElements* organizational unit. Each managed device e.g., ME1, ME2 contain a branch for storing operational logs (*opLogs*) and operational state (*opState*).

In a similar manner, all autonomic computing managers e.g., AM1, AM2 in the organization are placed under the *acManagers* branch of the *acElements* domain. From Figure 25, it can be seen that each manager has its own repository of policy objects that are applicable to the specific manager. A unit for storing operational information is also present.

As noted previously the *policyObjects* organizational unit consists of policy rules, conditions and actions. In Figure 25, under the policyRules branch are two rules named *PR0* and *PR1*. Each rule is identified by its *pcimRuleName* attribute. Recall from Section VII-A, that this is an attribute of the *pcelsRuleInstance* object class. Based on the above it can be inferred that all policy rules within the *policyObjects* branch of an autonomic manager must conform to the *pcelsRuleInstance* schema definition. Each policy condition under the *policyConditions* branch of Figure 25 must conform to the *pcimRuleConditionAssociation* object class definition, as its *pcimConditionGroupNumber* attribute is the basis for which policy conditions are identified (see Section VII-B). The same is true for the *policyAction* branch, only this time, any entry under this organizational unit must conform to the *pcimRuleActionAssociation* schema, as its *pcimActionName* attribute is the means by which policy actions are stored and accessed (see Section VII-C). Observe from Figure 24, that the *acManagers ou* also has a branch for policy objects i.e., policy rules, conditions and actions that have manager-wide applicability. In other words, all managers in the domain can share the policy objects under this branch.

Each manager also has a unit for entries relating to operational information (*opInfo*). In the DIT implemented in this project the *opInfo* consists of the state of the machine (opState) at points in time, as well as information relating to changes made to devices by the AM (*opLogs*). In Figure 26, the entries in operational logs (opLogs) and operational state (*opState*) organizational units are both identified by their individual UNIX time stamp (or Posix time).

Entries or information in a DIT are identified by their distinguished names or DN. For instance, the unique identifier for policy rule PR1 within the organizational structure is;

DN: pcimRuleName=PR1, ou=policyRules, ou=policyObjects, ou=AM1, ou=acManagers, dc=acElements, dc=acDomain

Fig. 27. Unique identifier for PR1

and the identifier for the operational state information of AM1 at (Posix) time 1301893100 is;

DN: timeStamp=1301893100, ou=opLogs, ou=opInfo, ou=AM1, ou=acManagers, dc=acElements, dc=acDomain

Fig. 28. Example identifier for AM opState information

B. Data Integrity and Security

In a directory that acts as a back-end for LDAP, data integrity is enforced through the twin mechanisms of object classes and attributes. Object classes are used to group attributes that apply to a specific component, and attributes contain data values for this component. The definition of the object class of a component indicates those attributes that are mandatory and those that are optional. The structures of the contained attributes in turn, specify the data types that values can take. For instance, values can be restricted to Integer, String or Boolean types etc., thus ensuring type safety. The syntax of attributes also indicates how attributes are to be compared during a search or compare operation. For example, an attribute that specifies a string value might indicate case sensitivity during a search etc. In addition, the attribute definition also indicates if it is multi-valued or single-valued. In Section VII, it was shown that object classes can include the attributes of other classes through inheritance or the use of auxiliary classes. This mechanism enables reusability and extensibility. As mentioned previously, all object classes and defined attributes must have globally unique identifiers assigned by IANA (see Section IV-A). All of these mechanisms help enforce entry integrity in the DIT. All related object classes and attributes are written to a schema file and the LDAP server is pointed to it. The syntax for object class and attribute definition are contained in RFC 2252.

Concerning data security, the LDAP allows for granting, restricting or denying access to any branch, attribute or distinguished name (DN) on a DIT using a user name/password mechanism. For instance, the domain controller of the DIT shown in Figure 24 might restrict access of one AM to a handful of managed element on the one hand, and grant full access to all managed element to another AM, on the other. In another scenario, the domain controller might give read access to its policy object entries to an AM but deny write access.

Apart from data integrity on the DIT, there is also a need to ensure integrity of the information exchanged between autonomic elements or between an autonomic element and an LDAP server located remotely. Recall from Section IV-D, that the interface between an autonomic element and its MIB or an external component i.e., *I-4*, only allows for LDAP-type transactions. Since LDAP uses TCP as its transport layer, it is able to ensure communication security by leveraging the Secure Sockets Layer (SSL) of the transport layer. If remote communication security is required, TCP port 636 is used, otherwise port 389 is used.

C. Data Availability through Referrals and Replication

LDAP allows branches of a DIT to be spread over several servers located at different physical locations. Regardless of the physical location of branches, the LDAP client e.g., an AM, still views the DIT as a consistent whole. The above is enabled by a mechanism known as Referrals. In a DIT where referral is implemented, rather than entries containing values, they would contain addresses to where the required data is housed. The object class for a referral entry is defined in RFC 3296 [34]. The LDAP client or server may resolve the referral. If the LDAP server is configured with *Chaining*, then the server gets the data from another server using the address contained in the instance of the referral object. Without *Chaining*, the referral is returned to the client and it is up to the client to issue a query based on the referral. Obviously, *Chaining* is the preferred method as it allows the whole process to be transparent to the client.

Whole copies of the DIT are allowed to be placed on multiple servers using a technique known as replication. Two types of replication configuration exist. The first, Master-Slave replication allows the Slave server to be updated by the Master. Client accessing the copy of the DIT on the Slave are only given read access. Write and update operations must be done on the Master server which then updates the Slave server after a defined period. If the server is configured without *Chaining* and a client attempts to write to the copy of the DIT on the Slave, the Slave server returns a referral for the Master server to the client. If configured with *Chaining*, the Slave server handles the update transparently. The second replication configuration is called Master-Master, which allows for reading, writing and updating on any of the LDAP servers. Changes to the DIT propagated to other servers later.

Both of these methods enable data availability, improve performance and ensure reliability. For instance, data can be placed closer to the consuming client through replication or referrals, thus reducing network overhead. In addition to the above, a backup of the DIT or branches of the DIT is always maintained through replication and referrals, respectively. Another benefit of these two mechanisms is that the DIT or parts thereof can be moved around possibly for scalability reasons without the need to change the client codes.

X. Mechanisms for Achieving Management Coordination in ACSs

The technical details of the IMD and ACS presented in Sections IV, V, VI, VII and IX provide the mechanisms by which the management coordination requirements set out in Section VIII are achieved. This section describes how each of these mechanisms or a combination of mechanisms is used to meet each requirement.

1) **Establishing administrative relationships:** The DIT structure shown in Figure 24 of Section IX-A is the basis for which administrative and security relationships are formed. In order to participate in an autonomic computing domain (*acDomain*), an AM must attempt to bind itself to the DIT of that domain. It does this by issuing an LDAP bind command through its I-4 interface to the *acDomain* (see Section IV-D). To place this joining request, the AM must have been configured with the right credentials i.e., username and password for the *acDomain* (see Section IX-B). If the bind request is successful, the *acDomain* creates a branch on the DIT for the new AM. Recall that the exact physical location of this new AM branch is irrelevant (see Section IX-C). As soon as the AM becomes aware of its new branch, it proceeds to set up its policy objects, operation and state information sub-branches. An ME is added to the domain much in the same way as an AM. All successful bind requests are recorded by the *acDomain*. This way, it is aware of all active objects within its sphere of influence. If an autonomic element no longer wishes to be part of the DIT, the element informs the domain controller of same.
2) **Resolving management conflict:** Management conflict can be resolved in two ways, once areas of potential conflicts have been identified. The first mechanism is known as hard resolution mechanism. Here, two or more AMs that may negatively interfere with one another are prevented from executing policy rules that point to the same *Policy Role Collection* Object (see Section VII-D). The soft resolution mechanism, which is the

second method, allows two or more AMs to use policy rules that point to the same *Policy Role Collection* Object but during periods where there is a risk of conflict, the ACS domain manager disables the policy rules. This is done by setting the *Enabled* attribute of the policy rule to *disabled* (see Section VII-A). Outside of the conflict risk period, the policy rule is enabled.

3) **Monitoring Autonomic Elements:** An autonomic element is able to persist current and previous state information in its *opState* branch on the DIT. It is also able to log its operational activities in the *opLogs* branch. *opLogs* and *opState* are depicted in Figure 26 for AM1 and in Figure 24 for ME1. If the state of an autonomic element needs to be verified, then it is simply a matter of querying its *opState* branch. This query will be based on the estimated entry time of the *opState* entry of interest. If the *acDomian* controller or a peer-AM requires information on why a managerial action was taken by an AM, a similar query with the time estimate is performed on the *opLogs* branch of the AM.
4) **Support for granting and requesting Services:** The *acDomain* controller supports and grants services to autonomic elements also using the DIT structure. For instance, assuming the proper administrative relationships have been established, an AM can query the *acDomian* for information relating to the available managed elements. Based on the retrieved information, the AM can then proceed to create its own policy objects for managing these elements. Requesting a bind to a DIT is also an example of support for services. Many other services specific to an *acDomain* can be defined, requested for and granted using the LDAP Interrogation, update and authentication and control operations (see Section IV-D).
5) **Reliable remote policy object communication:** Since LDAP relies on the TCP for network transport functionalities; an AM through its interfaces is able to reliably communicate policy actions or instructions to an ME and receive sensory information from the same ME. Keep in mind that TCP provides reliable ordered byte stream delivery to a network end device. SSL in TCP can also be used to provide security for autonomic elements when managerial transactions are carried out over a network (see Section IX-B).
6) **Policy object sharing:** An *acDomain* can define policy objects that are globally available to all AMs in its domain. For instance, in this project policy sharing is achieved by placing these common policy objects in the *policyObject* organizational unit of the *acManagers* branch (see Figure 24). Of course, the policy objects defined in the branch of an AM can be utilized by other AMs, if need be, assuming the right security relationships have been established. In the above example, one AM might be totally dependent on another AM 's policy object branch for its policy rules, conditions and actions. This may be a mechanism for enforcing hierarchy in a group of AMs. Recall from Sections VII-B and VII-C, that policy conditions and actions are associated with policy rules using their DNs. This mechanism allows two or more policy rules to reuse the same condition(s) or action(s), if necessary.
7) **A Policy rule Selection Mechanism:** The means to select the best policy rule for a particular *Context* is unique to the targeted application. Nevertheless, support for this exists in this work. Consider a scenario where two or more policy rules are applicable to a *Context*. An AM that conforms to the IMD architecture simply uses the *contextAmbiguity* and *contextResolved* message events defined in Section V to resolve the uncertainty.
8) **Low complexity and Reusability:** In this work, there are several levels of extensibility, which support low complexity and reusability. The reliance on standard based objects e.g., LDAP and its associated RFCs allow autonomic elements designed by different vendors to interact efficiently with one another, thus engendering low implementation complexity. If the *acDomain* becomes too large, it can be split into more manageable chunks without impacting on the structural integrity of the implemented DIT (see Section IX-C). This makes the *acDomian* scalable and by extension of low complexity. The security mechanisms utilized in this work are also well established. The policy object classes presented can also be extended in a structured manner to include attributes of vendor specific objects (see Sections VII-B and VII-C). In Section VI, suggestions were made regarding the implementation of functions that are self-contained and therefore reusable when attempts are made to extend the layer configuration of an AM.

XI. Conclusion

The technological reach of autonomic computing systems (ACS) is currently stymied by the lack of standardized certification procedures for these systems. This is made more difficult still by the lack of a consistent architecture for autonomic elements and systems and a deficit in standard metrics by which performances of these systems once built can be measured. In this paper, a structure based on already standardized protocols for the ACSs was proposed along with a flexible but consistent architecture for the autonomic manager (AM). Standard metrics are dealt with in the second part of this two-part paper.

Concerning the architecture for the AM, it was shown in this paper and elsewhere that due to implementation inconsistencies, the MAPE architecture was not well suited for certification purposes. As a result, an architecture that is flexible but guides the implementation more clearly than MAPE does was required. Biological animals, including humans use the same basic physiological structure to sense, process and effect changes in their immediate environment. Technically, this structure can be called a 'standard'. Incidentally, an architecture i.e., the Intelligent Machine Design (IMD) based on biological 'standard' had already been proposed. However, it lacked specific technical details to make it viable for ACSs or any other computing system for that matter. The first task in the process of expressing the three-layered IMD architecture for use in autonomic computing was to have it imbued with the four cardinal self-management properties i.e., self-configuration, self-healing, self-optimization and self-protecting. Each layer will implement these properties, albeit with differing levels of intelligence, computational complexities and execution speed, depending on both the architectural rules and on the application's requirements. The second task required the definition of four interfaces i.e., *I-1*, *I-2*, *I-3* and *I-4* for the IMD. Along with the definition of these interfaces, were the descriptions of the structure of the information communicated on each. Based on the make up of the IMD, five possible configurations, including their allowable message sequence charts were derived and presented. These configurations are one of the vehicles by which architectural flexibility is achieved. For the final task, the Policy Core Information Model (PCIM) framework was proposed as the basis for which policy rules, conditions, actions and repository are defined.

The Directory Information Tree (DIT) of the Lightweight Directory Access Protocol (LDAP) was proposed as the structure on which the

ACS is built. To support scalability in the system, branches of the DIT can be distributed over a number of physical locations and hardware, without affecting the consistency of the tree as viewed by elements. Apart from aiding scalability, distributing the branches of the tree also helps with data availability, as data can be transparently placed close to where it is consumed. This branch distribution is realized through a mechanism known as LDAP Referrals. Data availability can also be achieved by storing copies of whole or parts of the DIT in multiple locations. These copies are made consistent with the original using the LDAP server Replication mechanism. Replicating the DIT also achieves system robustness, as copies can be used as backups, if the main DIT becomes corrupted or unavailable.

Using the technical components of the DIT and the IMD, a number of issues relating to efficient management coordination and element interactions are resolved. These issues include but are not limited to; establishing security and administrative relationships, management conflict resolutions, autonomic element monitoring, support for extensibility and reusability across the system.

The proposals in this paper are foundational steps towards standardization of autonomic components, with a longer term goal of achieving certification of autonomic systems, which in turn is key to the long term acceptance and sustainability of the autonomic computing paradigm.

REFERENCES

[1] H. Shuaib, R. J. Anthony, and M. Pelc, "A framework for certifying autonomic computing systems," *The Seventh International Conference on Autonomic and Autonomous Systems: ICAS 2011*, pp. 122–127, May 2011.

[2] Autonomic Research Group, University of Greenwich *http://cms1.gre.ac.uk/research/autonomics/tech.html. Latest Access: December 20th, 2012*.

[3] M. Salehie and L. Tahvildari, "Autonomic computing: Emerging trends and open problems," *DEAS'05. Workshop on the Design and Evolution of Autonomic Application Software*, vol. 30, pp. 1–7, 2005.

[4] R. Sterritt, "Autonomic computing," *Innovations System Software Engineering (2005), Springer-Verlag*, vol. 1, no. 1, pp. 79–88, 2005.

[5] IBM, "An architectural blueprint for autonomic computing," *IBM Whitepaper*, June 2006.

[6] C. Reich, K. Bubendorfer, and R. Buyya, "An autonomic peer-to-peer architecture for hosting stateful web services," *CCGRID '08. 8th IEEE International Symposium on Cluster Computing and the Grid*, 2008.

[7] C. Dorn, D. Schall, and S. Dustdar, "A model and algorithm for self-adaptation in service-oriented systems," *ECOWS '09. Seventh IEEE European Conference on Web Services*, pp. 161 – 170, 2009.

[8] V. Nicolicin-Georgescu, H. B. R. Lehn, and V. Benatier, "An ontology-based autonomic system for improving data warehouses by cache allocation management," *Knowledge and Experience Management Workshop*, September 2009.

[9] J. Ferreira, J. Leitao, and L. Rodrigues, "A-osgi: A framework to support the construction of autonomic osgi-based applications," *Technical Report RT/33/2009*, May 2009.

[10] F. Mei, Y. Liu, H. Kang, and S. Zhang, "Policy-based autonomic mobile network resource management architecture," *ISNNS 10. Proceedings of the Second International Symposium on Networking and Network Security*, April 2010.

[11] B. Pickering, S. Robert, S. Mnoret, and E. Mengusoglu, "Model-driven management of complex systems," *MoDELS'08. Proceedings of the 3rd International Workshop on Models@Runtime , Toulouse, France*, pp. 277 – 286, October 2008.

[12] R. Quitadamo and F. Zambonelli, "Autonomic communication services: a new challenge for software agents," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 17, no. 3, pp. 457–475, 2008.

[13] B. A. Caprarescu and D. Petcu, "A self-organizing feedback loop for autonomic computing," *IEEE CS'09. In Proceedings of the Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns*, vol. 126–131, 2009.

[14] D. A. Norman, A. Ortony, and D. M. Russell, "Affect and machine design: Lessons for the development of autonomous machines," *IBM Systems Journal*, vol. 42, no. 1, pp. 38 – 44, 2003.

[15] R. Sterritt, M. Parashar, H. Tianfield, and R. Unland, "A concise introduction to autonomic computing," *Elsevier Journal on Advanced Engineering Informatics,*, pp. 181–187, 2005.

[16] M. Wahl, A. Coulbeck, T. Howes, S. Kille, Critical Angle Inc., Netscape Communications Corp., and Isode Limited, "Lightweight directory access protocol (v3): Attribute syntax definitions (RFC 2252)," December 1997.

[17] K. Zeilenga and OpenLDAP Foundation, "Lightweight directory access protocol (LDAP): Technical specification road map (RFC 4510)," June 2006.

[18] B. Moore, E. Ellesson, LongBoard-Inc., J. Strassner, A. Westerinen, and Cisco-Systems, "Policy core information model – version 1 specification (RFC 3060)," February 2001.

[19] B. Moore and IBM, "Policy core information model (PCIM) extensions (RFC 3460)," January 2003.

[20] M. Pana, MetaSolv, A. Reyes, Computer Architecture, UPC, A. Barba, D. Moron, Technical University of Catalonia, M.Brunner, and NEC., "Policy core extension lightweight directory access protocol schema (PCELS) (RFC 4104)," June 2005.

[21] J. Strassner, Intelliden Corporation, B. Moore, IBM Corporation, R. Moats, Lemur Networks, Inc., and E. Ellesson, "Policy core lightweight directory access protocol (LDAP) schema (RFC 3703)."

[22] M. Wang, N. Kandasamyt, A. Guezl, and M. Kam, "Adaptive performance control of computing systems via distributed cooperative control: Application to power management in computing clusters," *ICAC '06. IEEE International Conference on Autonomic Computing*, pp. 165–174, 2006.

[23] M. Zhao, J. Xu, and R. J. Figueiredo, "Towards autonomic grid data management with virtualized distributed file systems," *ICAC '06. IEEE International Conference on Autonomic Computing*, pp. 209–218, 2006.

[24] S. Ghanbari, G. Soundararajan, J. Chen, and C. Amza, "Adaptive learning of metric correlations for temperature-aware database provisioning," *ICAC '07. IEEE International Conference on Autonomic Computing*, 2007.

[25] B. Khargharia, S. Hariri, and M. S. Yousif, "Autonomic power and performance management for computing systems," *ICAC '06. IEEE International Conference on Computing*, pp. 145–154, 2006.

[26] J. Xu, M. Zhao, J. Fortes, and R. Carpenter, "On the use of fuzzy modeling in virtualized data center management," *ICAC '07. IEEE International Conference on Autonomic Computing*, 2007.

[27] R. Wang, D. M. Kusic, and N. Kandasamy, "A distributed control framework for performance management of virtualized computing environments," *ICAC '10. IEEE International Conference on Autonomic Computing*, pp. 89–98, 2010.

[28] M. Kutare, G. Eisenhauer, and C. Wang, "Monalytics: Online monitoring and analytics for managing large scale data centers," *ICAC '10. IEEE International Conference on Autonomic Computing*, pp. 141–150, 2010.

[29] X. Zhu, D. Young, B. J. Watson, Z. Wang, J. Rolia, S. Singhal, B. McKee, C. Hyser, D. Gmach, R. Gardner, T. Christian, and L. Cherkasova, "1000 islands: Integrated capacity and workload management for the next generation data center," *ICAC '08. IEEE International Conference on Autonomic Computing*, pp. 172–181, 2008.

[30] C. Kennedy, "Decentralised metacognition in context-aware autonomic systems: some key challenges," *In American Institute of Aeronautics and Astronautics (AIAA) AAAI-10 Workshop on Metacognition for Robust Social Systems, Atlanta, Georgia*, 2010.

[31] S. R. White, J. E. Hanson, I. Whalley, D. M. Chess, , and J. O. Kephart, "An architectural approach to autonomic computing," *ICAC '04. IEEE International Conference on Autonomic Computing*, 2004.

[32] M. Wahl and Critical Angle Inc., "A summary of the x.500(96) user schema for use with ldapv3 (RFC 2256)," December 1997.

[33] S. Kille, Isode Ltd., M. Wahl, Critical Angle Inc., A. Grimstad, R. Huber, and S. Sataluri, "Using domains in ldap/x.500 distinguished names (RFC 2247)," January 1998.

[34] K. Zeilenga and OpenLDAP Foundation, "Named subordinate references in lightweight directory access protocol (LDAP) directories (RFC 3296)," July 2002.

Towards Certifiable Autonomic Computing Systems Part II: Measuring and Rating Autonomic Computing Systems

Haffiz Shuaib and Richard John Anthony
Autonomics Research Group
School of Computing and Mathematical Sciences,The University of Greenwich.
Park Row, Greenwich, London SE10 9LS, UK
Email: haffiz.shuaib@yahoo.com, R.J.Anthony@gre.ac.uk

***Abstract*—Autonomic computing systems are a promising technology for bending the cost curve associated with information and communication technology (ICT) service management and for aiding the growth and evolution of complex computing systems. Indeed, this has motivated a significant amount of research. However, a central plank to achieving fully-fledged autonomic computing systems is missing i.e., the ability to certify these systems. The certification process will provide a basis; for assessing the quality of autonomic systems with similar functionalities, for assessing the current capability of the system and its suitability to the problem, to assess the impact of a certified component on a system and to resolve legal liability, if the autonomic computing systems were to fail.**

In this second part of a two-part paper, several steps to rate or certify autonomic computing systems within the context of the targeted application domain are proposed. In the first instance, the autonomic manager architecture proposed in the first part of this work is associated with indices that indicate how mature an autonomic machine is. The maturity index, the layer configuration of the machine and the implemented autonomic self-management properties are used to derive a mathematical expression that describes the machine in qualitative terms. These qualitative metrics in turn point to what quantitative measures or performance characteristics can be obtained from the machine under an evaluation scenario. The proposed quantitative metrics are based on the International Standard Organization's software quality specification i.e., ISO/IEC 9126. Using the software engineering standard for product evaluation i.e., ISO/IEC 14598-4, the four steps for certifying an autonomic computing system are outlined. Finally, an Ant Colony Optimization (ACO) application called Path Finder (PF) is used to demonstrate the proposals in this work.

Keywords-Autonomic computing systems; Certification; Performance; Verification; Measurement;

I. Introduction

A true Autonomic Computing System (ACS) is one that is able to automate the management decision-making process and reflect on the quality of the decisions made. This, it must do regardless of the environmental context and within the goals set by the human operator. The ultimate aim of autonomic computing systems is to allow complex Information Technology (IT) infrastructure to evolve to handle more difficult tasks or change in their immediate environment, without significantly increasing the cost of management.

As with most critical or increasingly complex systems, an ACS should and must be certified on the basis of its expected characteristics before it goes live, as these systems have applications from the financial to the space exploration industries. However, no progress has been made in the area of autonomic computing certification i.e., there is no framework to guide the process by which two or more autonomic machines are rated in relative terms, assuming these machines target the same application domain and no standard measure of performance for these systems [1]. A crucial aspect of correctly assessing the quality of an autonomic computing system is knowing what to measure and where to take these measurements. This task is often very difficult [2]. An attempt to address this difficulty is the primary objective of this paper. To this end, the following are proposed;

1) Qualitative measures that convey the complexity, intelligence and functionality of the autonomic machine.
2) Quantitative metrics that allow the autonomic machine to be measured based on specific performance attributes.
3) A fixed number of evaluation planning and execution steps that will lead to a final certification statement for an ACS.

These three objectives are dealt with within the context of the four cardinal self-management properties identified for ACSs i.e., self-configuration, self-optimization, self-protection and self-healing [3]. The idea is that since the four self-management properties to varying extents are applicable to all autonomic computing application domains, the certification framework can be made domain agnostic by defining it around these properties. This framework, once defined, is applied to an autonomic application use-case i.e., Path Finder (PF). The PF application is an Ant Colony Optimization (ACO) application in which an autonomic manager guides a managed object or robot along a gridded map. As with all ACO applications, the primary objective is to get the robot to the food source from the nest and back to the nest using the shortest route. This application is written in the C-Sharp programming language.

This paper collates together the findings of a detailed research project whose roadmap can be found in [1] and more extensive details in [4].

The rest of this paper is structured as follows; in the section following, a brief recap of the intelligent machine design (IMD) architecture covered in the first part of this paper is presented. In Section III, an expression that conveys a qualitative measure of an autonomic computing machine is derived. Quantitative metrics based on the normative framework of the International Standard Organization/International Electrotechnical Commission software quality specification [5] are discussed in Section IV. How these metrics apply to autonomic computing systems is presented in Section V. Section VI contains the steps for software evaluation based on ISO/IEC 14598 [6] that should lead to a final certification statement for an evaluated ACS. The proposals of Sections III - VI are demonstrated using the PF application in Sections VIII - XI. See Section XII for the conclusion. The appendix discusses special cases of the mathematical expression proposed in Section III.

II. The Intelligent Machine Design Architecture: A Recap

For completeness, a technical summary of the intelligent machine design (IMD) architecture is presented in this section. See Section II-B of the first part of this two-part paper [7] for a more exhaustive discussion on, and the philosophy behind this architecture.

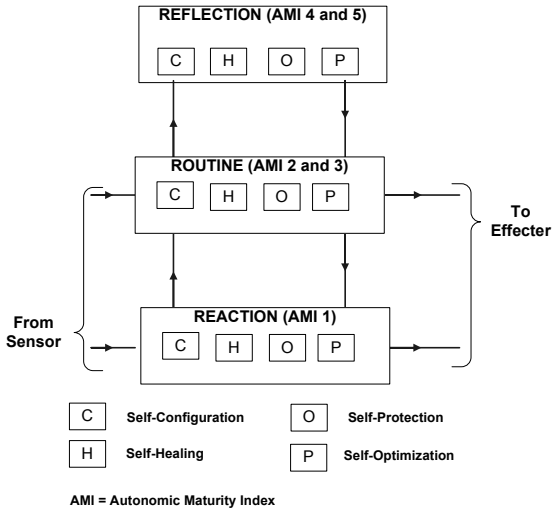


Fig. 1. An Autonomic Computing expression of the IMD (The AMI is discussed at length in Section III)

The IMD is made up of three distinct layers viz; The Reaction (R1), Routine (R2) and Reflection (R3) layers as shown in Figure 1. The Reaction layer is the least intelligent of all three, in that it accepts information from the sensory input and effects a change through its singular implemented policy rule. If this policy rule is unable to handle the input information to the machine, R1 passes control over to the Routine layer.

The Routine layer on the other hand, implements more than one policy rule and can select and apply the best rule from amongst these for the input context. As a result, the speed at which the Routine layer reacts to changes is expected to be relatively slower when compared to the Reaction layer. If the Routine layer is unable to find a suitable policy rule to effect a corresponding change or if two or more policy rules apply to an input context, it defers to the Reflection layer.

The algorithms implemented in the Reflection layer attempt either one of two things; (1) use of complex techniques e.g., artificial intelligence mechanisms (Fuzzy Logic etc.), to resolve policy rule conflicts flagged by the R2 layer or (2) create a new policy rule on the fly when none of the policy rules in the repository apply to the extant input context. The Reflection layer is the only one of the three that does not have sensory (S) inputs or effecter (E) outputs. The Reflection layer can inhibit or excite the activities of the Routine layer if required. The Routine layer can also do the same to the Reaction layer.

The three-layer structure of the IMD can thus be aligned with a simplified view of human problem solving in which there is a need to achieve rapid pre-programmed response in some situations, whilst being able to take a longer time to determine the best of several possible options in other situations, and ultimately being able to reason and learn new strategies in new situations. Using car driving as a means to explain this, consider emergency braking where the pre-programmed Reaction is faster than thinking speed; choosing which lane to enter at a fuel station, based on observing queue length (Routine); and studying the map to try to avoid an area of heavy congestion (Reflection).

III. Qualitative Metrics for Autonomic Computing Systems

It can be envisaged that autonomic computing products will offer a range of competing management services in the near future. If multiple systems are targeted at a specific application domain, then a means by which these systems are qualified from an autonomic perspective is required. Knowing the relative autonomic level at which a system operates will reveal what quantitative measurements can be extracted e.g., efficiency, latency etc. and what results to expect within the bounds of the complexity of a goal. This section contains the relevant discussions and proposals in this regard.

In the first subsection, the related works carried out in this area are discussed, together with apparent disadvantages. A five-level autonomic index that can be aligned with the IMD is proposed in Subsection III-B. These proposed indices are linked with the four cardinal self-management properties of autonomic computing systems in Subsection III-C. This link is to ensure that the proposed indices have relevance to most autonomic application domains.

A. Measuring Autonomicity

The term by which systems are ranked based on their particular autonomic capabilities go by a number of names in autonomic literature. For example, it is referred to as Autonomic Control Levels (ACL) in [2], Levels of Autonomy (LOA) in [8] and Degree of Autonomicity [9]. Still other papers term it the Autonomic Adoption Model (AAM) [3] or the Autonomic Computing Maturity Index (AMI) [10]. For consistency in this work, the preferred term will be AMI.

Several attempts have been made to describe criteria for which the AMI is to be based. For instance, [11] proposes using the following as the basis for assessing autonomic capabilities; the complexity of the objective, the operating environment and the level of human interaction with the machine. The motivation for the scale above was to have a consistent measure by which costs and suitability of a proposed robotic system for military operations can be ascertained.

In [12], a 10-point AMI that ranges from High (10) to Low (0) that depends on the relative influence of the participating entities i.e., Man or Machine on the following attributes was proposed;

- Information Acquisition: Reading, sorting, filtering and aggregating input data.
- Information Analysis: Performing complex computation on the acquired data e.g., prediction, data integration etc.
- Source of Decision: Making decisions based on the analyzed data. And
- Source of Action: Take an action based on the decision made.

The more the machine handles any of the above listed attributes, the higher up the scale the system is assessed to be.

An 11-point autonomic scale is presented in [2] based on similar attributes to [12]. The only difference is that they are labelled differently. The attributes of [12] i.e., Information Acquisition, Information Analysis, Source of Decision and Source of Action are called Observe, Orient, Decide and Act, respectively in [2].

The AMI proposed in [3] is characterized by what parts of the system's autonomic management activities are automated versus those that are manually implemented; The resulting five level autonomic scale is as delineated below;

- Manual Level: At this level all autonomic management activities are handled by the human operator.
- Instrument and monitor: Here, the autonomic system is responsible for the collection of information: This collected/aggregated information is analyzed by the human operator and guides future actions of the operator.
- Analysis Level: On this level, information is collected and analyzed by the system. This analyzed data is passed to the human administrator for further action(s).

- Closed loop Level: This works in the same way as the Analysis level, only this time the system's dependence on the human is minimized i.e., the system is allowed to action certain policies.
- Closed loop with business processes Level: At this level, the input of the administrator is restricted to creating and altering business policies and objectives. The system will operate independently using these objectives and policies as guides.

The AMI system proposed in [11] is targeted specifically at robots, thus limiting its application domain. At the time of its publication i.e., [11], its definition of the autonomic scale was a work in progress. The AMI in [12] is based on who makes the decisions and how these decisions are executed. Clough's AMI definition is also based on these two criteria [2]. From a certification perspective, the place of a system on the autonomic computing maturity index should be defined by who/what makes the decisions and the quality of the decisions themselves. After all, this is how human managers that autonomic systems are supposed to steadily replace would be evaluated. Both metrics will engender a certain level of trust in the system. The scales in [3] is said to be narrowly defined and technically vague [9]. This makes it difficult to align an autonomic system with these maturity indices [13]. These concerns do not help the certification process along. In the next subsection, an AMI that includes some of the advantages of the autonomic ranking system discussed in this section is proposed.

B. Autonomic Computing Maturity Indices (AMI)

The architecture shown in Figure 1 can be associated with the AMI. To do this, an attempt is made to expressly define what each Maturity Index means from a technical perspective, and further relate each index to the layers of the IMD. The Five maturity indices are thus interpreted as:

- **Maturity Index 1:** Here, only one policy action is executed in response to all input signals and encountered contexts. Complex operations are referred to the human operator or to the immediate higher layer. This maturity index corresponds to the Reaction layer.
- **Maturity Index 2:** This index corresponds to the Routine layer. If the Routine layer is unable to find a suitable policy rule from a policy repository or if there is a policy rule ambiguity, it relies on the human administrator to provide a new solution or resolve the policy rule conflict.
- **Maturity Index 3:** This is similar to Maturity Index 2, only that this time, the Routine layer consults the Reflection layer to solve its policy rule problems.
- **Maturity Index 4:** This index corresponds to the Reflection layer. The Reflection layer of a Machine in this index will attempt to solve the policy rule problem of the Routine layer, and monitor the implementation of this new policy rule. If the policy rule fails in its objective or if a new policy rule cannot be created, the human administrator is required to intervene.
- **Maturity Index 5:** This is similar to index 4, but rather than defer to the human administrator, if a suitable policy rule is not found or created, the algorithm within the Reflection layer will continually attempt to create a new policy rule or resolve the policy rule conflict. This index should be used to define autonomic machines that will be unable to get in touch with the human manager, a craft in deep space for example. Another possible example for this index is a scenario where the human intervention cannot be timely enough due to the complexities in the system.

In effect, the autonomic maturity level 1 corresponds to the Reaction layer, levels 2 and 3 correspond to the Routine layer, levels 4 and 5 correspond to the Reflection layer. The position of an autonomic computing system on the defined maturity indices above provides a possible basis for verifying the source of the decision making process and the quality of the decisions made. For instance, if a system in question specifies a Maturity Index of 2, the certification process would know that the 'court of last instance' is the human administrator. The certification process would now seek to verify the qualification of skilled personnel for the system to be awarded an index of 2. If the system seeks to be tagged with an index of 5 i.e., the decision making process is handled ultimately by the machine itself, the algorithm implemented in the Reflection layer must be shown to be robust enough to handle this task.

C. Autonomic Self-Management AMI Qualification

Regardless of the application domain targeted by an autonomic computing system, it is expected that some or all four of the self-management properties be implemented. The AMI proposed in the last subsection is able to achieve application domain agnosticism by being associated with these self-management properties, as opposed to tying it to a specific application.

Consequently, a mathematical expression that describes an autonomic manager that implements the IMD architecture is derived. The benefits of this expression include:

1) The verification of the characteristics of the autonomic manager at a glance. Examples of such characteristics are the complexity, implemented functionality of the machine, level of intelligence etc.
2) An indication of the extent to which a manager conforms to the IMD specification.
3) An indication of the self-management properties implemented in the manager.
4) Assisting the design of automated architectural verification algorithms of the implemented machine, if required.
5) It indicates the relevant quantitative metrics that ought to be measured. The certifier can use this as a guide during the evaluation process. Quantitative metrics are discussed in Section V.

Before deriving the expression describing the machine, it is instructive to restate here that only five layer configurations are allowed for the IMD. Section VI of Part I has shown why Expressions (1) - (5) are the only legal IMD configurations, and as such the explanation here is kept brief. All IMD configurations must contain either R1 or R2, as these layers are the only ones connected to the sensory and effecter mechanisms. If the R3 layer is implemented then the R2 layer must also be present, as R3 cannot communicate with R1 directly.

$$R1 \leftrightarrows \emptyset \leftrightarrows \emptyset \tag{1}$$

$$R1 \leftrightarrows R2 \leftrightarrows \emptyset \tag{2}$$

$$R1 \leftrightarrows R2 \leftrightarrows R3 \tag{3}$$

$$\emptyset \leftrightarrows R2 \leftrightarrows \emptyset \tag{4}$$

$$\emptyset \leftrightarrows R2 \leftrightarrows R3 \quad (5)$$

Where R1, R2 and R3 represent the Reaction, Routine and Reflection layers, respectively and $\leftrightarrows$ represents a connection between the layers of the machine.

Observe from Figure 1 that all four self-management properties can be implemented in all layers of the IMD, although with varying degrees of intelligence, speed and complexity. Expression (6) is used to symbolically represent the self-management properties implemented by a specific layer of the IMD.

$$SM_{RX} = \{SC_X, SH_X, SO_X, SP_X\} \quad (6)$$

where $X = \{1, 2, 3\}$ is the level of the IMD where the relevant self-management property is implemented. Thus SC_X, SH_X, SO_X, SP_X represent the implementation of the self-configuration, self-healing, self-optimization and self-protecting properties at layer X, respectively.

If all four self-management properties are implemented in the Routine layer (R2) then symbolically it is represented by $SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\}$. Similarly, if the Reaction layer (R1) does not implement the self-configuration property then the appropriate representation is $SM_{R1} = \{\emptyset, SH_1, SO_1, SP_1, \}$. In other words, if a self-management property is not implemented, its corresponding symbol is replaced with $\emptyset$ in the enclosed set of Expression (6).

Based on the discussion so far, an IMD compatible Autonomic Manager can be described by a combination of its AMI, IMD layer configuration and the implemented self-management properties as shown in Expression (7).

$$AM_z = \left\{ \begin{array}{l} AMI; R1 \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, S_{O1}, SP_1\} \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\} \end{array} \right\} \quad (7)$$

where $AMI = \{1...5\}$.

Consider the AMs described by Expressions (8), (9), (10) and (11).

$$AM_1 = \left\{ \begin{array}{l} 4; R1 \leftrightarrows R2 \leftrightarrows \emptyset; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\} \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\} \end{array} \right\} \quad (8)$$

$$AM_2 = \left\{ \begin{array}{l} 4; \emptyset \leftrightarrows \emptyset \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\} \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\} \end{array} \right\} \quad (9)$$

$$AM_3 = \left\{ \begin{array}{l} 4; R1 \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\} \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\} \end{array} \right\} \quad (10)$$

$$AM_4 = \left\{ \begin{array}{l} 2; R1 \leftrightarrows R2 \leftrightarrows \emptyset; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\} \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{\emptyset, \emptyset, \emptyset, \emptyset\} \end{array} \right\} \quad (11)$$

AM_1 in Expression (8) is invalid because it specifies an AMI value of 4 but does not implement the Reflection layer (R3). Recall from the last subsection that AMI 4 and 5 reside at the Reflection layer. Since R3 is not implemented in AM_1, it should specify $SM_{R3} = \{\emptyset, \emptyset, \emptyset, \emptyset\}$ not $SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\}$. AM_2 (see Expression 9) is invalid since the layer configuration $\emptyset \leftrightarrows \emptyset \leftrightarrows R3$ is not among those allowed for the IMD (see Expressions 1-5 at the beginning of this section). AM_3 and AM_4 in Expressions (10) and (11) conform to the IMD rules and thus valid.

If all four AMs were to go through a certification process, then AM_1 and AM_2 would have failed the first test and resources need not be expended to verify them or measure other attributes further. If AM_3 and AM_4 were designed in a way that both targeted the same application domain, from Expressions (10) and (11), one can tell that AM_3 will be expected to adapt to contexts that deviate from the norm, as it implements a Reflection layer (R3). Tests to verify this superior management capability should be carried out on AM_3 but not on AM_4, as it does not have an R3. The AMI of AM_3 also points to the fact that it is more mature in relative autonomic terms. These are just a few examples of how an expression describing an AM can point to what can and cannot be measured quantitatively (see the next section for discussions on quantitative metrics).

There are certain AM implementations that can appear to reveal some inconsistencies when attempts are made to describe them using Expression (7). These apparent inconsistencies are treated in the appendix.

IV. Quantitative Metrics for Software Evaluation

The Qualitative metrics presented in the last section while relevant to the certification process, lack a means by which an autonomic system is measured on a scale of magnitudes. A number of quantitative metrics must be derived to address the above. The International Standard Organization software quality specification i.e., ISO/IEC 9126-1998 is the basis on which the autonomic quantitative measures presented in the next section are defined.

ISO/IEC 9126 [5] defines six main characteristics that can be used to assess the quality of a software product, including; Functionality, Usability, Portability, Reliability, Efficiency and Maintainability. These normative characteristics and their attributes can be used to pose certain questions to the evaluation of an autonomic computing machine. The answers to these questions, which may be boolean or numerical ratios can be used to derive a single value or set of values that form the basis for which a system to be certified is rated.

In terms of the ***Functionality*** characteristic the following questions are posed:

1) *Suitability*: Does a function exist within the implemented system that provides for a specifically stated or implied need?
2) *Accuracy*: If it does, how well does it meet that need?
3) *Interoperability*: Is it able to interact with other systems e.g., AMs deployed in the same environment?
4) *Security*: How well does it prevent unauthorized access to the system data?

Questions relating to the ***Reliability*** characteristic include:

1) *Maturity*: What is the Mean Time to Failure of the system?
2) *Fault Tolerance*: Can the system maintain a specific level of performance in the face of a fault? i.e., how robust is the system?
3) *Recoverability*: Can the system regain peak performance after the impact of a failed component is mitigated?

The ***Usability*** characteristic deals with how user-friendly the system is. This characteristic is not discussed further in this work because it is subjective and directly dependent on how the system is developed and the system's targeted application domain.

The attributes of the ***Efficiency*** characteristic ask the following questions:

1) *Time (temporal) behaviour* : How much time does it take to complete a task or does the system meet the hard or soft execution time constraints?
2) *Resource Utilization*: How much resources in terms of memory space, CPU cycles and network bandwidth are committed to achieving the task?

The ***Maintainability*** characteristic deals with the following questions:

1) *Analysability*: How well are system faults and their causes recognized and understood?
2) *Changeability*: Can the system or part of it be modified easily?
3) *Stability* : When the system is modified, how well does it perform thereafter?
4) *Testability* :Regardless of whether changes are made can the system be validated?
5) *Modularity and Coupling*: Can the system be expressed in specific component parts and can these parts be joined together efficiently?

Finally, the questions associated with the Portability characteristic deals with:

1) *Adaptability*: Can the software be adapted for another environment using only components contained within?
2) *Installability*: How easy is it to install?
3) *Co-Existence*: When installed in a new environment can it co-exist with other installed components?
4) *Replaceability*:Can it efficiently replace another software designed for the same purpose?

A *Compliance* attribute applies to all six characteristics discussed above. This attribute seeks the answer to the following question; How well does the autonomic machine conform to specified standards/conventions etc.?

These six software quality characteristics are applicable to two types of metric namely ***Internal*** and ***External***. The Internal metrics are applicable to the quality of the actual code, while the External metrics apply to the operational behaviour of the software code. While both metrics are equally important, evaluation of the Internal metric should be left to the autonomic computing system developer. The ACS certifier should only be concerned with the External metric for the following reasons:

1) The developer of the code might not want to reveal the internal logic, thereby protecting intellectual property rights and trade secrets. This must not pose a barrier to certification.
2) The operational external behaviour of the code under a rigorous test is sufficient to inform on whether the autonomic machine does what it says it does.
3) Points 1 and 2 allow the containers of the code e.g., the three layers of the IMD to act as black boxes that can be tested, thereby lessening the amount of work on the certifying authority without compromising the quality of the certification process.

V. Quantitative Metrics for Autonomic Computing Systems

The broad ISO/IEC software evaluation characteristics discussed in the last section are applied to autonomic computing systems in this section. Specifically, the methods for computing these quantitative metrics, the outputs of these computations and the interpretation of these outputs are proposed and presented.

A. Functionality

There are several dimensions to measuring the functionality characteristic in this work. In the first, the autonomic machine or system is measured in terms of the self-management properties it implements. The second dimension of this metric involves the level at which these management properties reside e.g., R1, R2 or R3. Note that the functionality behaviour expected at each of the three levels of the IMD will differ. For instance, at the R3 level, the assessor will want to verify if the Reflection layer is able to create a new policy or resolve a policy conflict on behalf of the Routine layer (R2). At the R2 level, the assessor will be looking at how well the Routine layer engages the Reflection layer when a context deviates from the norm, how well does the implemented policy selection algorithm execute given a specific context and how well it regulates the behaviour of the Reaction layer (R1). R1 will be evaluated based on how well it executes its implemented policy and how well it is able to engage R2 when the extant context cannot be dealt with.

From the above the following tasks and measurement types are derived based on the attributes of the functionality metric:

1) Suitability:

- **R1 Functional Suitability**
  Task: Find out if the self-management property i.e., self-configuration, self-healing, self-optimization or self-protection is supposed to be and is implemented in R1. Note that this can be deduced from the expression describing the machine, specifically SM_{R1} in Expression (7).

  Output Metric: 1 for Yes and 0 for No for each of the applicable self-management properties.

- **R2 Functional Suitability**
  Task: Find out if the self-management property i.e., self-configuration, self-healing, self-optimization or self-protection is supposed to be and is implemented in R2. Note that this can be deduced from the expression describing the machine, specifically SM_{R2} in Expression (7).

  Output Metric: 1 for Yes and 0 for No for each of the applicable self-management properties.

- **R3 Functional Suitability**
  Task: Find out if the self-management property i.e., self-configuration, self-healing, self-optimization or self-protection is supposed to be and is implemented in R3. Note that this can be deduced from the expression describing the machine, specifically SM_{R3} in Expression (7).

  Output Metric: 1 for Yes and 0 for No for each of the applicable self-management properties.

2) Accuracy: A functional self-management property that evaluates to 'Yes' after the suitability check is tested a number of times to verify its functional accuracy when it executes. In order to compute the measure of accuracy (A) the total number of times the self-management property is tested is counted and assigned to a variable N_{total}. The number of tries in N_{total} for which it executed as expected are counted and assigned to the variable $N_{success}$. Correspondingly, N_{fail} holds the total number of times the self-property fails the task. The accuracy of the task is given as the ratio of $N_{success}$ and N_{total} i.e., $A = \frac{N_{success}}{N_{total}}$. Functional elements that evaluate to 'No' at the functional suitability stage are awarded

a value of 0. This implies that the range of values for accuracy is $0.0 \leq A \leq 1.0$

- **R1 Functional Accuracy**
  Task 1: What is the accuracy with which the singular policy of R1 executes?

  Output Metric 1: $A_1 = \frac{N_{success}}{N_{total}}$

  Metric 1 Interpretation: The closer the value of A_1 is to 1.0 the better.

  Task 2: How accurate is the logic at R1 that passes control from R1 to R2, if the context cannot be properly handled by R1?

  Output Metric 2: $A_2 = \frac{N_{success}}{N_{total}}$

  Metric 2 Interpretation: The closer the value of A_2 is to 1.0 the better.

- **R2 Functional Accuracy**
  Task 1: Count the number of times R2 succeeds ($N_{success}$) in selecting the best policy for the specific context after N_{total} number of tries.

  Output Metric 1: $A_1 = \frac{N_{success}}{N_{total}}$

  Metric 1 Interpretation: The closer the value of A_1 is to 1.0 the better.

  Task 2: Count the number of times R2 correctly engages R3 or the human operator to solve policy conflicts or request the creation of a new policy.

  Output Metric 2: $A_2 = \frac{N_{success}}{N_{total}}$

  Metric 2 Interpretation: The closer the value of A_2 is to 1.0 the better.

  Task 3: Count the number of times R2 correctly moderates the actions of R1 by overwriting the policy executed by R1.

  Output Metric 3: $A_3 = \frac{N_{success}}{N_{total}}$

  Metric 3 Interpretation: The closer the value of A_3 is to 1.0 the better.

- **R3 Functional Accuracy**
  Task 1: Verify how accurate R3 is at creating a new policy or resolving policy conflicts reported by R2.

  Output Metric 1: $A_1 = \frac{N_{success}}{N_{total}}$

  Metric 1 Interpretation: The closer the value of A_1 is to 1.0 the better.

  Task 2: How accurate is R3 when signalling to the human operator that it is unable to solve a problem reported by R2. This task assumes that the machine operates at AMI level 4 (see Section III-B).

  Output Metric 2: $A_2 = \frac{N_{success}}{N_{total}}$

  Metric 2 Interpretation: The closer the value of A_2 is to 1.0 the better.

3) Interoperability: In Section X of the first part of this paper, a number of interoperability mechanisms where proposed to aid management coordination. These mechanisms included proposals for; (1) establishing administrative relationships, (2) resolving management conflict, (3) monitoring autonomic elements, (4) support for granting and requesting services, (5) policy sharing and (6) reliable remote policy communication.

Task: Let $O_{success}$ be the number of interoperability mechanisms implemented and O_{total} be the total number of interoperability mechanisms required for the application under consideration.

Output Metric: $I = \frac{O_{success}}{O_{total}}$

Note that $O_{total} \leq 6$.

Metric Interpretation: The closer the value of O is to 1.0 the better.

4) Security: In Section X of Part I of this work, a mechanism for establishing administrative and security relationships between elements of an ACS was presented. These relationships are established through the *I-4* interfaces of the IMD. The IMD has four different connections that use the *I-4* interface (see Figure 3 in Part I). This Security attribute verifies that all the active *I-4* interfaces are secure, as per the standard security procedure laid out for establishing security relationships.

Task: Count the total number of active *I-4* interfaces that conform to the security conventions and standards and store this value in $I_{success}$. Let I_{total} be the total number of active I-4 interfaces.

Output Metric: $S = \frac{I_{success}}{I_{total}}$

where $I_{total} \leq 4$.

Metric Interpretation: The closer the value of S is to 1.0 the better.

B. Reliability

Notice from Section IV, that the maturity, fault tolerance and recoverability attributes of the Reliability characteristic correspond to the self-management properties of self-healing and self-protection. Since the presence and accuracy of these self-management properties are already dealt with in Section V-A, there is no need to further define metrics for them here.

C. Usability

The usability metric is not discussed within the context of this work for the reasons given in Section IV.

D. Efficiency

The Efficiency attributes of Time behaviour and Resource utilization are measured during the period in which the functional accuracy tests are carried out. Note that the only relevant values for efficiency are those obtained when the measured accuracy tallies with the expected outcome i.e., $N_{success}$.

1) Time Behavior: ***Task***: This is a measure of how much time it takes to complete a self-management task i.e., T_{SM}. The time behaviour will naturally be impacted by memory access times, computation duration and network delays.

Output Metric:

$$T_{avg} = \frac{\Sigma^{N_{success}} T_{SM}}{N_{success}}$$

where T_{avg} is the average latency of an executed self-management property and $N_{success}$ the number of times for which the executed self-management property behaves as expected. The T_{avg}value can be linked to real-time managerial constraints in autonomic computing applications.

Metric Interpretation: The lower the value of T_{avg} the better.

2) Resource Utilization: ***Task***: Measure the amount of resources, specifically network bandwidth consumed (R_{ntw}), the amount of memory (R_{mem}) and the CPU utilization (R_{cpu}) used to achieve the self-management task. This metric enables the ACS or AM to be tested for compliance with the resources available on a given platform.

Output Metric: R_{ntw}, R_{mem} and R_{cpu} in bits/second, bytes and %, respectively.

Metric Interpretation: Generally, the lower the values of R_{ntw}, R_{mem} and R_{cpu}, the better.

E. Maintainability

The Maintainability characteristic is associated primarily with the architecture implemented, in this case the IMD reference architecture. As such the attributes for the Maintainability characteristic is discussed within the context of the IMD.

1) Analyzability: Since this attribute has to do with reporting failure or operational anomalies, the reporting agent must be able to precisely describe what the problem is and alert the agent in charge of addressing anomalies. Recall from Section IX of Part I that an autonomic element is able to write its undertakings to an operational log (see Figure 26 in Part I). Tabs can be placed on an autonomic element and corresponding actions taken by monitoring its operational logs. This, of course, assumes that the right administrative relationships have been established.

Task:Verify that faults injected into the system are properly logged and responded to for an autonomic element. Let MA_{total} be the total number of injected faults and $MA_{success}$ the total number of times the faults are resolved.

The analyzability metric is denoted as MA.

Output Metric:

$$MA = \frac{MA_{success}}{MA_{total}}$$

where $0 \leq MA \leq 1$

Metric Interpretation: The closer the value of MA is to 1.0 the better.

2) Coupling and Modularity: Within the context of this work, coupling and modularity (C & M) has to do with how the layers are arranged or configured based on the need of the targeted application domain. The Expressions for the five valid machine configurations are given in Section III-C. If these configurations are violated, then an autonomic manager fails the modularity and coupling test.

Task: Verify that the implemented configuration is valid.

Output Metric: 1 for Yes and 0 for No.

3) Stability: Assuming a layer of a machine implementing the IMD as its reference architecture was to go out of commission, without violating the permitted layer configurations, would other layers of the machine operate as usual.If the answer to this question is yes, the machine is deemed to be stable.

Task: Decommission 1 or 2 layers of the machine without violating the valid layer configurations, to see if the remaining layer or layers operate as expected.

Output Metric: 1 for Yes and 0 for No.

4) Testability: If the relevant functional characteristics of the ISO 9126-1998 specification can be applied to the machine under test, then it is testable. The output metric for this attribute is 1 for Yes and 0 for No.

F. Portability

1) Adaptability: This attribute relates to the ability of the autonomic machine to be adapted from one operating environment to another without modification. For example, if the autonomic system has been written using a programming language that runs on a virtual machine available to a number of considered hardware/software platforms, then the autonomic system is said to be adaptable. Adaptability is also aided in the proposed scheme by the fact that the IMD uses already standardized protocols, including the Policy Core Information Model (PCIM) [14][15] and the Lightweight Directory Access Protocol (LDAP) [16].

Task: Count the number of hardware/software platforms being considered. Store this value in P_{total}. Identify the number of platforms on which the autonomic system will run without modification. Store this value in P_{run}.

Output Metric: $P = \frac{P_{run}}{P_{total}}$

Metric Interpretation: The closer the value of P is to 1.0 the better.

2) Co-Existence: This is a measure of the impact a deployed ACS has on other systems running within the same resource domain. Note that the generation of this metric, and its interpretation, are highly system dependent.

Task: Count the number of applications that are impacted negatively by a deployed ACS and set this value as C_{total}.

Output Metric: C_{total}

Metric Interpretation: The lower the value of C_{total}, the better.

3) Installability and Replaceability: These attributes are not treated in this work.

VI. Certifying Autonomic Computing Systems

The International Organization for Standardization/ International Electrotechnical Commission defines software evaluation as a four-step process in ISO/IEC 14598 [6]. These four steps include; Establishing the evaluation requirements, Specifying the evaluation, Designing the evaluation and Executing the evaluation. These four steps are discussed here and subsequently applied to the evaluation of an application use-case later in this paper.

A. Establish the evaluation requirements

There are three tasks associated with this step. The first of these tasks has to do with establishing the purpose of the evaluation. This will involve a description of the autonomic computing application, specifically its goals. The second task is to identify the type of software product to be evaluated. In other words, is this a complete product or a component of a larger software application or a product still undergoing development? Specifying the quality characteristics that are to be measured is the last task for this step. Note that these characteristics are the six defined in ISO/IEC 9126 discussed in Section IV.

B. Specify the evaluation

This step also consists of three tasks including: (1) Selecting the quality metric type i.e., External or Internal (see last paragraph of Section IV). This task also involves establishing a common procedure for assigning values for measured attributes. The same measurements under the same conditions should produce similar and consistent values. (2) Establishing a rating level for each of the selected metric. This task mandates the certifier to state what specific quantifiable values for the measured quality characteristics are acceptable and those that are not. For example, the functional accuracy for a self-management property acceptable for one application might be 0.9 and for another 0.5 depending on the targeted application (see Section V-A2 for the interpretation of the functional accuracy metric). (3) Assigning weights to certain measurements. For instance, if the targeted application domain is one where computing resources are in abundance but operations need to be completed within a tight interval, then it is sensible that attributes relating to resource utilization should be assigned a lower priority and those relating to time behaviour assigned a relatively higher weight.

C. Design the evaluation

The design of the evaluation procedure will depend on the autonomic application. For instance, the evaluation for an application targeted at space exploration will almost certainly be carried out within a simulated environment. Other applications may be amenable to real time testing in the field. Regardless, the evaluation plan and design must be such that the most significant environmental test factors are considered. The plans and designs for evaluating an autonomic application is done in this step.

D. Execute the evaluation

This activity involves measuring the relevant quality characteristics identified in Section VI-A based on the evaluation plan set out. The computed results are matched against the acceptability rating criteria established in Section VI-B. With the ratings, a final verdict is pronounced on the system. For example, three possible certification statements are:

The system meets all the specified and implied needs of the end user.
or
The system meets all the specified and implied needs of the end user but requires more hardware to meet the acceptable efficiency characteristic rating.
or
The system does not adhere to relevant specifications and conventions and therefore does not meet the specific and implied need of the end user.

VII. The Application Use-Case

In Part I of this paper, an application called Path Finder (PF) was used to demonstrate some of the technical proposals made. For consistency, PF is used to demonstrate how the IMD and the certification process proposed previously can be applied to an autonomic application.

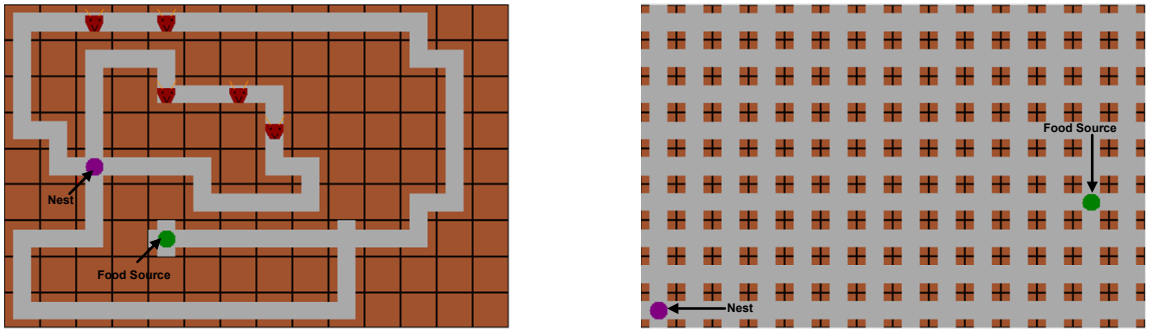
The PF which is an Ant Colony Optimization (ACO) application was selected as a use-case not only because it was sufficiently complex to demonstrate the proposed technical mechanisms but also because ACOs have a wide variety of applications. For instance, it is relevant to several domains, including but not limited to robotics, engineering, computer networks, finance, resource and job scheduling etc. Specifically, in this section, a relatively detailed description of the application is given. Its architectural configuration with respect to the IMD and its implemented policy framework is defined.

A. Application Description

The PF application in which robots are guided to and fro between a base (nest) and a target (food source) is used as a means of demonstrating the process of evaluation and comparison of AMs, using the techniques described in this work. A number of AMs are devised to navigate a robot in a maze from its base, to a target and back again in a simulated environment. The AMs have differing sophistication, using a variety of techniques to find the best route. The evaluation involves measuring a number of aspects of the AM's performance but the purpose of the exercise is not to find the absolute performance of these AMs, rather, the emphasis here is on showing how the evaluation of the AMs is performed and how they can be rated in terms of their suitability for purpose, accuracy and efficiency.

In PF, a robot begins from the nest and tries to find its way to the food source on a gridded map. When the food source is found, the robot must then navigate its way back to the nest. This process is repeated for the duration of the experiment. Regardless of how many times the robot has found the food source, when it gets to the nest it forgets the position of the food source and begins afresh to locate it. The reason for this is to mimic varying food source locations. Each robot has only local knowledge and only stigmergic communication occurs between robots i.e., robots are only able to influence one another indirectly through pheromones left on the paths. The lack of global knowledge requires that robots depend on intelligent search and navigational algorithms to find the food source or the nest. Each robot is controlled by a separate AM instance. In autonomic parlance, the robot is the Managed Resource while the AM is the Manager Element.

At each time step or clock tick, an AM must decide and then move a robot to the next square on the deployed map. The maps considered in this work are shown in Figures 2(a) and 2(b). The AM is allowed to move the robot in one of four available directions i.e., Top, Bottom, Left or Right square. The ability of a manager to steer its robot efficiently from the nest to the food source and back is linked to the level of Intelligence of the search algorithm within the manager. The implemented algorithms are discussed in detail in Section VIII-A. The position of the food source and the nest can be



(a) Closed Map (b) Open Map

Fig. 2. Closed and Open Maps

seen on both maps. It is worth mentioning that the closed map in Figure 2(a) is less complicated to navigate than that shown in Figure 2(b). The reason for this is that the open map has more pathways, thus creating relatively more opportunities for the robot to wander or get lost when searching for a target. Within the context of this work, a round trip from the nest to the food source and back to the nest is known as a ***Home Run***. It should be noted that the collective behaviour of the deployed robots and their managers as it relates to the objective is what is important, not the performance of the individual robots.

B. Application Use-Case Architectural Design

Recall from Figure 1, that all layers of the IMD are able to implement the four self-management properties. The evaluation focuses on the self-optimization management property i.e., it is the duty of the self-optimization mechanism of the implemented IMD to attempt to find the optimal path between the nest and the food source targets and move the robot accordingly. In Section VII-D3, it is shown that moves to the Top, Bottom, Right and Left Square on the gridded map are realized by four different policy rules. From the maps shown in Figure 2, on any clock tick, at least two moves (of the four valid moves) are possible. Recall from Section V of Part I, that two or more valid rules for a Context always generates a *ContextAmbiguity* event message which must then be handled by the R3 layer of the AM. Since every Context will generate a *ContextAmbiguity* event message, the R1 layer of the IMD is not required. As a result, Configuration V of the IMD, shown in Figure 3 and its message sequence shown in Figure 4, is the most appropriate for the PF application (see Section VI of Part I).

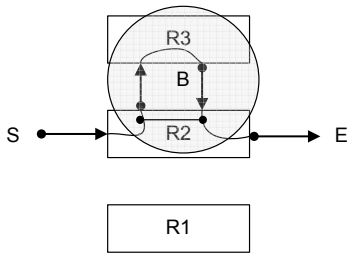


Fig. 3. Configuration IV ($\emptyset \leftrightarrows R2 \leftrightarrows \emptyset$)

With the description given above, the AM structure for the PF application can be described by Expression (12).

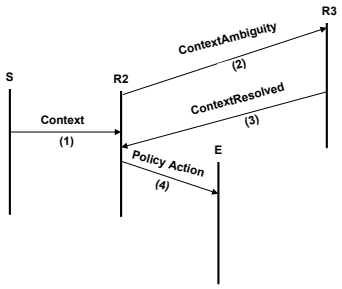


Fig. 4. Message Sequence for Configuration V

$$AM = \left\{ \begin{array}{l} 5; \emptyset \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{\emptyset, \emptyset, \emptyset, \emptyset\}, SM_{R2} = \{\emptyset, \emptyset, SO_2, \emptyset\} \\ SM_{R3} = \{\emptyset, \emptyset, SO_3, \emptyset\} \end{array} \right\} \quad (12)$$

The general form of the Expression can be found in Section III-C i.e., Expression 7. From the expression it can be seen that since R1 is omitted from the *PF* application, all self-management properties of R1 i.e., SM_{R1} are set to null. For R2 and R3 only the self-optimization property is implemented for the reasons given earlier. The self-configuration, self-healing and self-protection properties are set to null for R2 and R3, as can be seen in SM_{R2} and SM_{R3} in Expression (12). The numeral 5 in the Expression means that the machine operates at the Autonomic Maturity Index (AMI) of five (5) i.e., there is no human involvement in the operation of the autonomic managers for this application.

C. Application I-1 and I-2 Interface Information Structure

In Section IV of Part I, it was said that the information for the Sensory (S) interface, I-1 and that for the Effecter (E) interface, i.e., I-2 must conform to the object class structure stipulated in RFC 2252 [17]. The object classes: *pfSensory* and *pfEffecter* defined in Sections IV-A and IV-B of Part I of this paper are used for the I-1 and I-2 interfaces of the *PF* AM, respectively.

D. Application Use-Case Policy Framework Design

The Policy rules, conditions and actions to be used by any AM targeting this application domain are to follow the standardized policy core information model (PCIM) framework specified in RFC 3060 and 3460. The compliant rules, conditions and actions that should be implemented are defined in the subsections that follow.

1) Policy Condition Format: The defined PCIM policy condition class for this application i.e., ***pfCondition*** has six attributes, namely:*isActive*, *isValidMove*, *topDirection*, *bottomDirection*, *rightDirection* and *leftDirection*. All these attributes are of Boolean data types and are discussed in Section VII-B of Part I of this paper.

$$C = \{isActive, isValidMove, topDirection, bottomDirection, rightDirection, leftDirection\} \quad (13)$$

then the structures of the four realizable policy conditions are;

$$Policy\ Conditions = \begin{cases} C1 = \{True, True, True, False, False, False\} \\ C2 = \{True, True, False, True, False, False\} \\ C3 = \{True, True, False, True, False, False\} \\ C4 = \{True, True, False, False, False, True\} \end{cases}$$

An explanation of how a policy condition of type *pfCondition* object class is evaluated can be found in Section VII of Part I.

2) Policy Action Format: The policy action implemented in the *PF* application is of type *pfAction* and it consists of a single object attribute i.e., *functionID*. As noted in VII-C of Part I, it is of data type string and it points to the function that creates an instance of the *pfEffecter* object class.

If *A* in Equation (14) is the general structure of a policy action for this application;

$$A = \{functionID\} \quad (14)$$

then the four acceptable instances of the *pfAction* class are;

$$Policy\ Actions = \begin{cases} A1 = \{moveTop\} \\ A2 = \{moveBottom\} \\ A3 = \{moveLeft\} \\ A4 = \{moveRight\} \end{cases}$$

3) Policy Rule Format: Recall from Part I, that a policy rule consists of one or more policy conditions and actions and attributes that govern how these conditions are evaluated and how actions are to be executed. Conditions and actions are associated with a policy rule by adding the distinguished names (*DNs*) of the relevant policy conditions and actions to the rule's *ConditionList* and *ActionList* attribute, respectively. If the condition within a policy rule evaluates to true, all actions within that policy rule are executed subject to the other attributes of the rule. For the *Path Finder* application considered in this report, only four policy rules are required. These rules are based on the four conditions and actions described in the last two subsections.

Assume that Equation (15) represents the general structure of a rule;

$$R = \{C, A\} \quad (15)$$

where *C* is the *DN* of a policy condition and *A* the *DN* of a policy action.

The valid policy rules for the *PF* are;

$$Policy\ Rules = \begin{cases} R1 = \{C1, A1\} \\ R2 = \{C2, A2\} \\ R3 = \{C3, A3\} \\ R4 = \{C4, A4\} \end{cases}$$

VIII. Evaluation Requirements for Use-Case

This is the first of four defined steps set out by ISO /IEC 14598 that must be carried out when attempting to evaluate a software system (see Section VI-A). There are three tasks associated with this activity. These tasks are discussed within the context of the *PF* application in Sections VIII-A- VIII-C.

A. Purpose of Evaluation

The purpose of this evaluation is to establish how well in relative terms six different autonomic managers (AM) are able to achieve the objectives of the *pathfinder* (*PF*) application described in Section VII-A. All six AMs considered in this work implement a Fuzzy Logic algorithm in the R3 layer. This algorithm helps the AM decide the next move for its robot. Note that placing the Fuzzy Logic algorithm in this layer is in line with the idea that the R3 layer is the most intelligent of all three layers and houses the artificial intelligence mechanism, if implemented. The difference between the six evaluated AMs (described later) is in their implemented path search method and the quality of the inputs to the Fuzzy Logic algorithm.

Before discussing the AMs , it is pertinent to mention the common features of the managed robots:

- When a robot has found food, it deposits pheromones on the squares of the map it transverses on its way back to the nest. Depending on the navigational algorithm implemented in the AM, these pheromones may or may not be used by other robots as inputs to the Fuzzy Logic algorithm. Deposited pheromones evaporate after a controlled number of ticks of the clock.
- The robots do not have a global view of the considered map.
- Every time a robot sets out from the nest, it has no idea where the food source is located. The idea is to mimic multiple food sources.

The distinguishing characteristics of the evaluated AMs are presented below:

1) **Autonomic Manger-Basic (AM-B)**: In order to locate the food source or the nest, this AM uses an algorithm called **Basic**. In the **Basic algorithm**, the AM remembers the position of the last four squares crossed by its robot leading up to its current position. The current location of the robot, the available squares for the next move, the information relating to the last four squares crossed and the pheromones detected in each of these squares are passed to the *R3* layer by the *R2* layer of the manager using a *contextAmbiguity* message. The Fuzzy logic algorithm in *R3* uses this information as input and then outputs what it considers the best move for the robot. This output is sent to R2 using a *contextResolved* message. *R2* selects the policy rule that corresponds to that move and passes the associated policy action to the Effecter (E). The effecter moves the robot accordingly. The message sequence for this process is shown in Figure 4. Note that the Basic algorithm does not consider the relative strength of the pheromones on the squares when deciding the next move.
2) **Autonomic Manger-Pheromones (AM-P)**: This AM implements an algorithm similar to AM-B, only this time the square with the strongest pheromones are given higher weights when the decision is being made for the next move towards the

food source. The algorithm implemented by AM-P is called **Pheromones**.

3) **Autonomic Manger-Memory(AM-M)**: AM-Memory implements a variant of the Basic algorithm called the **Memory algorithm**. The difference here is that rather than remembering the last four squares traversed, the AM remembers all squares its robot passed through before getting to the food source. Once the food source is found, it traces its way back to the nest using the same path i.e., the one stored in its memory
4) **Autonomic Manger-Hill Climbing 1 (AM-HC1)**: This AM implements an algorithm called the **Hill Climbing 1 algorithm.** Like the Memory algorithm, it remembers the squares crossed to get to the food source, but unlike the Memory algorithm, it does not follow the route in its memory blindly back to the nest. It uses the stored information more selectively by applying the Hill Climbing search algorithm. The output of the Hill Climbing algorithm is used as one of the inputs to the Fuzzy Logic algorithm when the robot is trying to find its way home. This AM does not use pheromones.
5) **Autonomic Manger-Hill Climbing 2 (AM-HC2)**: AM-Hill Climbing 2 implements the **Hill Climbing 2 algorithm**. This algorithm is similar to that implemented in AM-HC1. The difference is that AM-HC2 additionally uses pheromones in the environment as inputs to its Fuzzy logic algorithm in exactly the same way as the Basic algorithm used in AM-B.
6) **Autonomic Manger-Hill Climbing 3 (AM-HC3)**: Implemented in this AM is an algorithm called **Hill Climbing 3**. This utilizes the Hill Climbing algorithm to find the shortest path back to the nest while giving higher weights to trails with the strongest pheromones when selecting the path to the food source in the same manner as AM-P.

B. Type of Product

The autonomic managers targeted at this application are self-contained and part of a larger system. They do not require additional components to carry out their activities.

C. Product Quality Model

The software quantitative metrics evaluated are:

- The Functionality attributes of *Suitability* to the objective and Accuracy with which the objective is achieved, if achieved at all (see Sections V-A1 and V-A2).
- Two attributes of the Maintainability characteristics are dealt with in this evaluation i.e., the Coupling and Modularity (C & M) attributes (see Section V-E2).

IX. Evaluation Specification for Use-Case

There are three tasks associated with this second evaluation activity as discussed in Section VI-B. These tasks are applied to the *PF* application and are presented in Sections IX-A- IX-C.

A. Metrics Selection

From the description of the PF application in Section VII, it is clear that the *R2* layer is mostly dependent on the *R3* layer for the policy rule choice for the next move of the robot. Put more succinctly; the machine is heavily dependent on the *R3* layer. As a result, only the Functional *Suitability* and *Accuracy* of R3 as it relates to the machine's self-optimization property is evaluated. Note that it is only when the Functional *Suitability* is confirmed to be a 1 is the *R3* Functional *accuracy* computed. All six AMs evaluated are functionally *suitable*.

The procedure for computing the Functional *Accuracy* metric of a self-management property at the *R3* layer is given in Section V-A2 as;

$$A_1 = \frac{N_{success}}{N_{total}} \tag{16}$$

where N_{total} is the total number of times the *R3* layer was called upon to resolve the a policy rule conflict/ambiguity by the *R2* layer and $N_{success}$ the total number of times these *R2* requests were successfully resolved.

Within the context of the *PF* application, $N_{success}$ is interpreted as the number of home runs (N_{hr}) achieved within the time considered. N_{total} represents the total number of steps i.e., the total number of squares (C_{sqr}) crossed by the robots, as instructed by the navigational algorithm in R3 to achieve N_{hr}. N_{hr} and C_{sqr} cannot be plugged into A1 in Equation (16) directly as these are two different quantities. A means to convert N_{hr} to steps or squares is required. To do this, the lowest (ideal) number of squares that can be traversed to achieve a single home run on a map is found and this number is the value of a single home run in squares (C_{hr}) for that map. For instance, for the Closed map in Figure 2(a), the lowest number of squares between the nest and the food source is 21 squares. Therefore, *42* (2×21) squares must be crossed for a round trip or 1 home run i.e., $nest \mapsto food \mapsto nest$ in the best case scenario. This implies that for the Closed map $C_{hr} = 42$ squares. For the Open map in Figure 2(b), $C_{hr} = 30$ squares. With the above, the R3 functional accuracy for the closed and open map can now be computed using Equation (17);

$$A_1 = \frac{N_{success}}{N_{total}} = \frac{C_{hr} * N_{hr}}{C_{sqr}} \tag{17}$$

The other metric evaluated in this work are the Coupling and Modularity attributes of the Maintainability characteristic (see Section V-E2). For the *PF* application, if the AM conforms to the configuration shown in Figure 3, the metric for these attributes is assigned a Yes or *1*, otherwise it is assigned a *0*. All six AMs evaluated conform to the configuration.

B. Rating Levels

A Functional *Suitability* metric value of 1(Yes) is a must for all AMs considered. For the purpose of this evaluation, if an AM is able to achieve a value of R3 Functional accuracy greater or equal to 0.4 for the self-optimization property, it is accepted. A Coupling and modularity value of 1(Yes) is mandatory for all evaluated AMs.

C. Criteria for Assessment

All three attributes discussed in Section IX-A to be measured i.e., *Suitability*; *Accuracy*; *Coupling and Modularity* are to be assigned equal weights in the final certification.

X. Evaluation Design for Use-Case

To assess the relative capabilities of the AMs with respect to the objective i.e., maximizing the number of home runs, each type of AM is tested in a simulated environment. The AMs guide their respective robots through the Closed and Open maps shown in Figures 2(a) and 2(b) for a set amount of time. Each simulation is run 30 times. At the end of each set of 30 simulation runs, the mean number of home runs (N_{hr}) is computed and extracted at a 95% confidence interval. Also extracted from the experiments are the number of squares traversed (C_{sqr}) to achieve the home run (N_{hr}) values. The C_{sqr} and the mean ordinate of the N_{hr} values, in conjunction with the C_{hr} value of the map under consideration are subsequently used to compute

the Functional *Accuracy* (A1). This is done using Equation (17). The Functional *Accuracy* along with the other two relevant metrics are compared to the threshold set in Section IX-B to arrive at a final certification statement for each type of AM.

It should be noted that pheromones deposited on the squares of the map are set to evaporate after 10 ticks for the evaluation carried out.

As mentioned in Section VII-A, the metrics evaluated are considered based on the collective behaviour of AMs of a similar type, not on individual AM performance.

XI. Evaluation Execution for Use-Case

The execution of the evaluation process defined in Sections VIII-X is repeated for two scenarios. In the first set of evaluations, the execution run time is fixed and the number of robots is varied for each type of AM. For the second set of execution, the simulation run time is varied, while the number of deployed robots is fixed.

A. Evaluation I

As noted previously, a robot is moved to the next square *en route* to the food source or nest at the tick of the clock. In this first evaluation, the number of ticks is held constant at 1200 ticks for each robot on both the Closed and Open maps. The number of robots deployed is steadily increased from 20 to 100 in intervals of 20. Since it is the collective behaviour of the robots that is being evaluated, if 20 robots are deployed in a scenario then there will be 24000 moves in total i.e., 20 robots $\times$ 1200 ticks. Table I shows the cumulative number of squares crossed (C_{sqr}) or moves by the robots based on their deployed numbers for both maps.

TABLE I
Cumulative Number of Squares (C_{sqr}) traversed by robots on the Closed and Open map

No. of deployed robots	Total No. of Moves
20	24000
40	48000
60	72000
80	96000
100	120000

From the graph shown in Figure 5, it can be seen that robots controlled by managers of types AM-P, AM-HC2 and AM-HC3 have the best performance in terms of the number of Home Runs (N_{hr}) achieved on the Closed map. The relative superior performance of these three AMs has to do with the quality of the mechanism by which the food source is found and the algorithm that guides the robot back to the nest. For instance, for the Closed map, managers of type AM-P and AM-HC3 have an advantage when it comes to finding the food source, as their navigational algorithms are biased towards paths with the strongest Pheromones. Type AM-HC2 and AM-HC3 managers are better at finding the nest on the return trip, primarily because both rely on the *Hill Climbing* algorithm to achieve this. Although, managers of type AM-HC1 also employ the *Hill Climbing* algorithm when trying to find the nest, its ability to achieve a comparative number of home runs when compared to AM-HC2 and AM-HC3 is hampered by the fact that it does not rely on Pheromones to find the food source.

The poor performance of managers of type AM-M is due to the fact that if its robots take the non-optimal route to find the food source, they will also take the non-optimal route back to the nest, as dictated by the **Memory** algorithm. The inability of AM-B to compete favourably with the other AM types is due to the mechanism

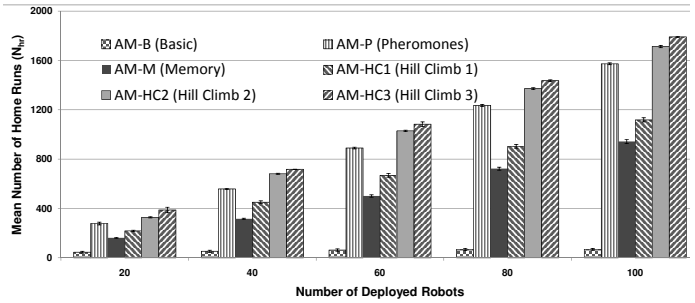


Fig. 5. Home Runs (N_{hr}) for the Closed Map

of the implemented **Basic** algorithm. Recall that managers with this algorithm have short-term memory with respect to where the robots have been. Another disadvantage is that the algorithm does not give higher weights to paths with the strongest pheromones. As expected for most of the AM types save AM-B, an increase in the number of deployed robots increases the number of home runs achieved.

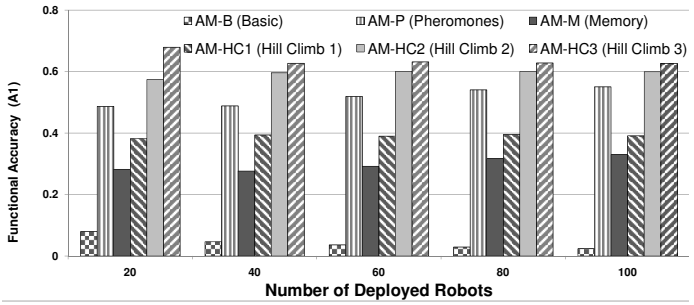


Fig. 6. Functional *Accuracy* (A1) for Closed Map

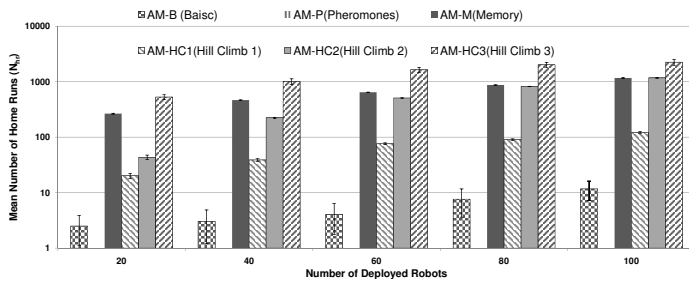


Fig. 7. Home Runs (N_{hr}) for the Open Map

The Functional *Accuracy* (A1) of the 6 types of AMs is computed and shown in Figure 6. These values of A1 are computed using Equation (17) with the mean ordinates of the Home Runs of Figure 5 as N_{hr}, the corresponding values of C_{sqr} in Table I and the C_{hr} value for the Closed map as inputs.

From the Figure and for the Closed map, it can be seen that only AM-P, AM-HC2 and AM-HC3 meet the 0.4 threshold set for R3 Functional *Accuracy* in Section IX-B.

The performances of the AMs on the Open map tell a slightly different story from the Closed map. The most significant difference is in the performance of the managers of type AM-P. In Figure 7, the number of home runs (N_{hr}) achieved by the AMs implementing the **Pheromones** algorithm is Zero. Recall that the Open map is a more complicated map, in that robots wander about; as there are many more paths here than those on the Closed map. A close examination of how

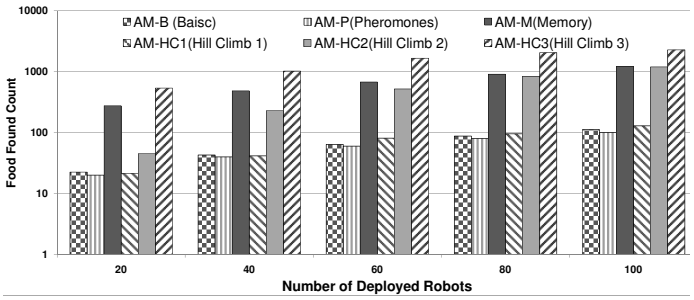


Fig. 8. Food Found Count for Open Map

the **Pheromones** navigational algorithm works revealed the reason for this subpar performance. Figure 8 shows the amount of times the food source was found by the various navigational algorithms. Notice that the AMs with the **Pheromones** algorithm found the food source a number of times. This suggests that the problem has to do with the robots not being able to find their way to the nest afterwards. This problem is further compounded by the fact that given the increased opportunity to roam on the Open map, robots with food are unable to find their way home will deposit pheromones on every square crossed. This in turn will lead other robots looking for the food source astray, as type AM-P managers depend heavily on paths with the strongest pheromones. Note that type AM-HC3 managers also have a bias towards paths with stronger pheromones but unlike the AM-P, the *Hill Climbing* algorithm implemented in managers of type AM-HC3 intelligently guides robots back to the nest. Hence, the reason for the higher home run counts for AM-HC3 shown in the figure. Another significant change of note is the performance of robots controlled by type AM-M managers. It appears the ability of the Memory algorithm to trace the path back to the nest, albeit suboptimally, was an advantage in terms of home runs achieved relative to the **Pheromones** algorithm as shown in Figure 7.

Again, the performance of robots managed by AM-B was relatively poor for the same reasons given for its performance in Figure 5. However, the non-reliance on the path with the strongest pheromones by the **Basic** algorithm is the reason for its relative higher performance in terms of home runs achieved when compared to those of managers of type AM-P.

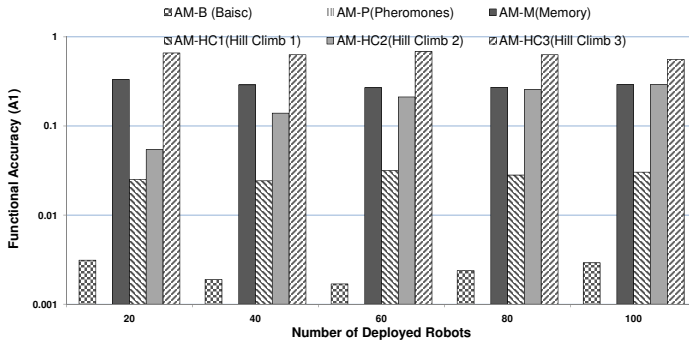


Fig. 9. Functional *Accuracy* for the Open Map

The Functional *Accuracy* (A1) for the Open map in this evaluation is shown in Figure 9. The computation of A1 was done in the same manner as that shown in Figure 6. Note that the $C_{sqr} = 30$ for the Open map (see Section IX-A). From Figure 9, it can be seen that only managers of type AM-HC3 meet the required threshold of 0.4 for Functional *Accuracy* set in Section IX-B.

B. Evaluation II

For the evaluation contained in this section, the number of robots was held constant at 20 and the number of clock ticks increased steadily from 24000 to 72000 in steps of 12000. As the robots are allowed to cover more ground (or squares) given the increased number of ticks, one would expect that the number of home runs achieved would increase as well. This is the case as shown in Figure 10 and 12 for the Closed and Open map, respectively. Even though the parameter varied in this evaluation is different from that varied in the evaluation of Section XI-A, the analysis of Evaluation I applies here as well. This can be verified from the graphs depicted in Figures 10 - 14.

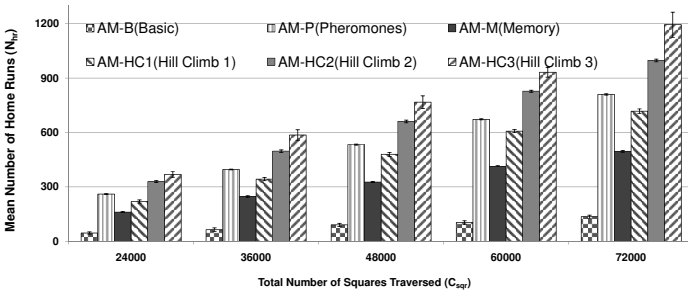


Fig. 10. Home Runs (N_{hr}) for the Closed Map

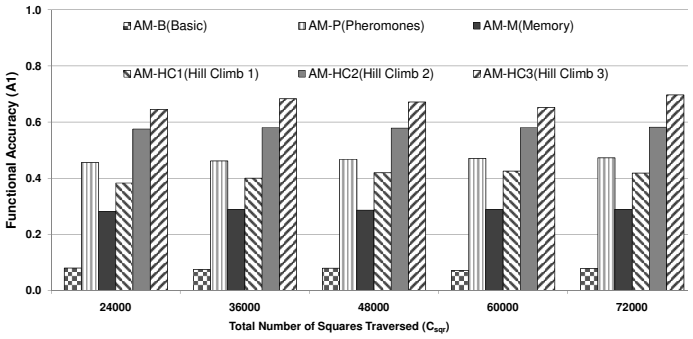


Fig. 11. Functional *Accuracy* for Closed Map

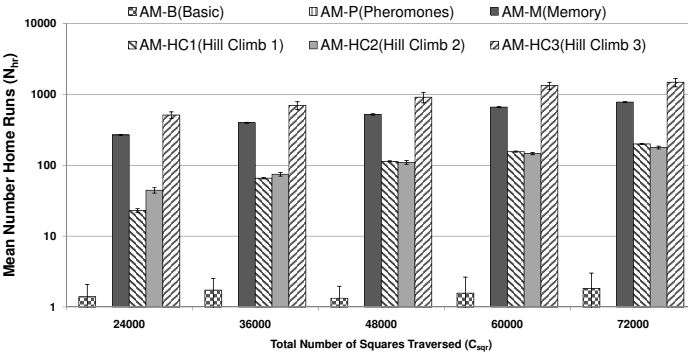


Fig. 12. Home Runs (N_{hr}) for the Open Map

From Figure 11, it can be seen that for the Closed map AM-P, AM-CH2 and AM-CH3 consistently meet the Functional *Accuracy* threshold set in Section IX-B i.e., 0.4. Managers of type AM-M meet

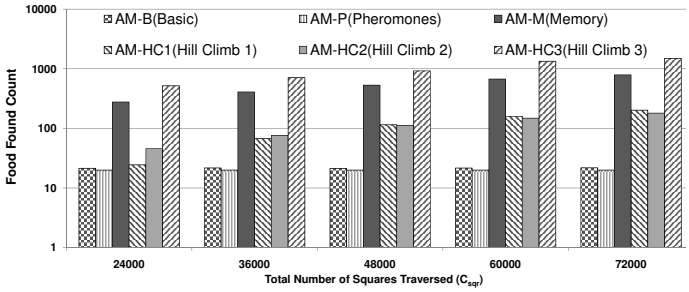


Fig. 13. Food Found Count for Open Map

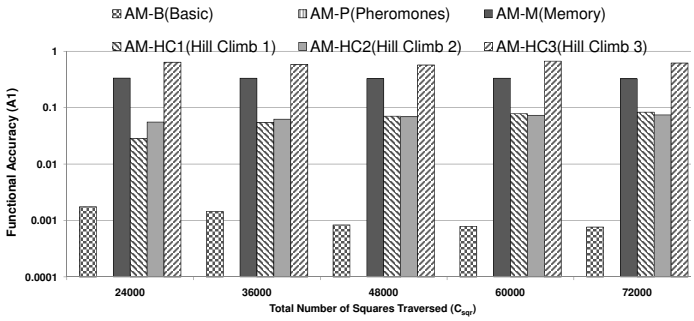


Fig. 14. Functional *Accuracy* for Open Map

this threshold only when the number of steps exceeded 24000. For the Open map, only managers of type AM-CH3 meet this requirement (see Figure 14).

C. Evaluated Autonomic Manager Acceptability

In Section IX, three metrics were identified as the basis for which an AM targeted at the PF application is accepted. These metrics were; Functional *Suitability* (FS), *R3* Functional *Accuracy* (A1) and Coupling and Modularity (C & M). The thresholds for acceptability for these three metrics were set in Section IX-B. All AMs targeting the PF application must meet the FS requirement (i.e., a 'Yes' value). For the A1 metric, all AMs must achieve a value of 0.4 or better in terms of the objective of the PF. The C & M, like the FS must have a value of 1 or a Yes.

TABLE II
MANAGER RATINGS FOR EVALUATION I (NUMBER OF DEPLOYED ROBOTS VARIED ON THE CLOSED MAP

AM Type	C & M	FS	A1 (Closed map)
AM-B(Basic)	✓	✓	
AM-P(Pheromones)	✓	✓	✓
AM-M (Memory)	✓	✓	
AM-HC1 (Hill Climb 1)	✓	✓	
AM-HC2 (Hill Climb 2)	✓	✓	✓
AM-HC3 (Hill Climb 3)	✓	✓	✓

From Section IX-A, it was made clear that all the AMs evaluated followed the architectural configuration of Figure 3, and as such, the C & M metrics is achieved (i.e., 'Yes') for all of them. The AMs were also said to be functionally suitable to the objective of the PF (see Section IX-A), therefore they all meet the FS attribute. The different performances of the AMs as it relates to the Functional *Accuracy* were presented in Sections XI-A and XI-B. Based on this, Tables II, III, IV and V show how each type of AM measured up to expectations with regards to the selected metrics for Evaluation I and II. Where a metric is met, the column for that metric is ticked for an AM type, and where it falls short, the column is left blank.

TABLE III
MANAGER RATINGS FOR EVALUATION I (NUMBER OF DEPLOYED ROBOTS VARIED ON THE OPEN MAP

AM Type	C & M	FS	A1 (Open map)
AM-B(Basic)	✓	✓	
AM-P(Pheromones)	✓	✓	
AM-M (Memory)	✓	✓	
AM-HC1 (Hill Climb 1)	✓	✓	
AM-HC2 (Hill Climb 2)	✓	✓	
AM-HC3 (Hill Climb 3)	✓	✓	✓

TABLE IV
MANAGER RATINGS FOR EVALUATION II (NUMBER OF CLOCK TICKS VARIED ON THE CLOSED MAP

AM Type	C & M	FS	A1 (Closed map)
AM-B (Basic)	✓	✓	
AM-P (Pheromones)	✓	✓	✓
AM-M (Memory)	✓	✓	
AM-HC1 (Hill Climb 1)	✓	✓	
AM-HC2 (Hill Climb 2)	✓	✓	✓
AM-HC3 (Hill Climb 3)	✓	✓	✓

TABLE V
MANAGER RATINGS FOR EVALUATION II (NUMBER OF CLOCK TICKS VARIED ON THE OPEN MAP

AM Type	C & M	FS	A1 (Open map)
AM-B (Basic)	✓	✓	
AM-P (Pheromones)	✓	✓	
AM-M (Memory)	✓	✓	
AM-HC1 (Hill Climb 1)	✓	✓	
AM-HC2 (Hill Climb 2)	✓	✓	
AM-HC3 (Hill Climb 3)	✓	✓	✓

Based on the contents of Tables II, III, IV and V, it can be said that only managers of type AM-HC3 consistently meet the rating levels set by the PF certifiers.

XII. CONCLUSION

To address the lack of a framework for certifying autonomic computing systems (ACSs) a novel five level Autonomic Maturity Index (AMI) was proposed and applied to the Intelligent Machine Design (IMD) architecture. These defined indices reflect how independent the AM is i.e, the amount of resources the AM is expected to expend in terms of intelligence, computational complexity and speed before it engages the human operator to solve a management task. An Expression that describes an autonomic machine was derived. The parameters for this expression include the AMI, the layer configuration of the machine and the implemented self-management properties. As a consequence, the derived expression indicates if the machine conforms to the dictates of the IMD architecture or not. If it does not conform, no further certification activity is carried out on the machine. If the machine's expression satisfies the established architectural rules, the certification process continues by measuring the machine's attributes relating to performance using the proposed quantitative metrics. How these metrics are computed and how the results are interpreted was discussed. These quantitative metrics are based on the six software quality characteristics of the ISO/IEC 9126-1998 specification i.e., the Functionality, Usability, Portability, Reliability, Efficiency and Maintainability characteristics. Evaluation steps that utilize the proposed quantitative and qualitative metrics for autonomic computing certification purposes were presented. These steps were

guided by the ISO/IEC 14598 software evaluation specification.

To demonstrate applicability of the architectural and metric systems proposed for building and certifying Autonomic Computing Systems (ACSs), an Ant Colony Optimization application called Path Finder (PF) was chosen. The main goal of managers targeted at the *PF* application is to guide robots from the nest to a food source and back. Six different managers were designed and implemented, each with a different navigational algorithm. The purpose was to compare the performances of these AMs to a specified rating level. From an architectural perspective, the *PF* demanded that each manager conform to Configuration V of the Intelligent Machine Design (IMD). Since the objective of the *PF* application is to find the most optimal route between the nest and the food source, only the self-optimization property of the four autonomic self-management attributes is implemented in the machine.

With respect to certification, all four procedures leading up to a final certification statement on the AMs were followed. Three metrics were identified as relevant to the evaluation of AMs targeted at the PF application. These metrics include; Functional *Suitability*, Functional *Accuracy* and *Coupling and Modularity*. Given the conformity of each type of AM to Configuration V of the IMD, each type of AM scored full marks for the *Coupling and Modularity* metric. They were also awarded full marks for Functional *Suitability*. For Functional *Accuracy*, each type of AM was evaluated and rated within the context of how well they performed when directing their robots on two different maps. Based on these three metrics, it was shown that only managers of type AM-HC3 met the specified rating threshold consistently.

APPENDIX

This appendix deals with a couple of special cases that may arise when trying to define an autonomic manager using Expression (7) presented in Section III-C.

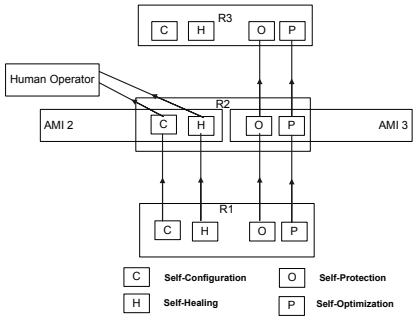


Fig. 15. Autonomic Manager 1 (AM_1) Before Normalization

Consider a situation where an AM is designed such that some of the self-management properties specify a particular AMI while others specify another. For instance, the designers of AM_1 shown in Figure 15, specify an AMI of 2 for the self-configuration and the self-healing properties in R2. The self-optimization and self-protection properties specify an AM_1 of 3. Since this machine specifies two AMIs it cannot be described using Expression 7 (see Subsection III-C).

$$AM_1 = \left\{ \begin{array}{l} ????; R1 \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, SO_1.SP_1\}, \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{\emptyset, \emptyset, SO_3, SP_3\} \end{array} \right\} \quad (18)$$

To rectify this anomaly, the machine shown in Figure 15 and partially described by Expression (18) must be normalized. The

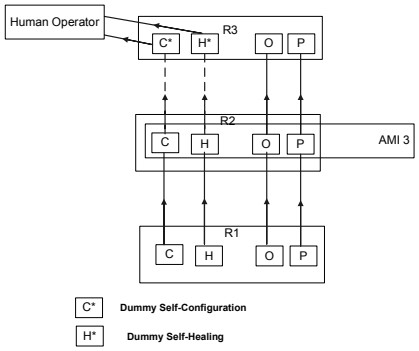


Fig. 16. Autonomic Manager 1 (AM_1) After Normalization

normalization process for AM_1 involves creating dummy implementations of the self-configuration and the self-protection functions in R3. These dummy implementations are empty functions that simply redirect any request from R2 to the human operator as shown in Figure 16. The normalized AM design is rated with the higher of the two AM_1 values it specified before normalization. The normalized expression for AM_1 is shown in (19).

$$AM_1 = \left\{ \begin{array}{l} 3; R1 \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\}, \\ SM_{R2} = \{SC_2, SH_2, SP_2, SP_2\} \\ SM_{R3} = \{SC_3^*, SH_3^*, SO_3, SP_3\} \end{array} \right\} \quad (19)$$

The asterisks in the SM_{R3} variable of Expression (19) signify dummy function implementation. A comparison between Figure 15 and 16 shows that the normalization process does not violate the intention of the AM designers.

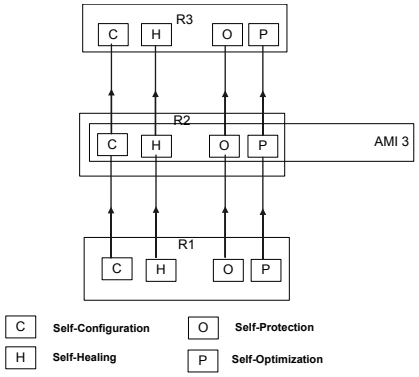


Fig. 17. Autonomic Manager 2 (AM_2)

$$AM_2 = \left\{ \begin{array}{l} 3; R1 \leftrightarrows R2 \leftrightarrows R3; \\ SM_{R1} = \{SC_1, SH_1, SO_1, SP_1\}, \\ SM_{R2} = \{SC_2, SH_2, SO_2, SP_2\} \\ SM_{R3} = \{SC_3, SH_3, SO_3, SP_3\} \end{array} \right\} \quad (20)$$

Recall that from the AMI of an expression describing an AM, one can quickly deduce a number of characteristics including how intelligent and complex the system being observed is. Speaking strictly from an AMI perspective, expressions (19) and (20) erroneously convey a sense of equivalence between AM_1 in Figure 16 and AM_2 in Figure 17. In fact, since AM_1 is able to rely on the human operator for some of its management function in R2; it should be judged lower than AM_2. In a case like this, some of the

quantitative metrics discussed in Section V will ensure that AM_2 is rated relatively superior to AM_1 at the end of the certification process. For instance, when the self-management functions at R3 of AM_1 are being measured for the functionality accuracy metric(A), the dummy functions will be rated with a zero value (see Subsection V-A2). This will not be the case for AM_2; thus ensuring its higher overall rating.

ACKNOWLEDGMENT

The authors would like to extend their appreciation to the anonymous contributor aptly named *pseudonym67* [18] for the navigational algorithm library and the user interface used in the evaluation aspect of this work.

REFERENCES

[1] H. Shuaib, R. J. Anthony, and M. Pelc, "A framework for certifying autonomic computing systems," *The Seventh International Conference on Autonomic and Autonomous Systems: ICAS 2011*, pp. 122–127, May 2011.

[2] B. T. Clough, "Metrics, schmetrics! how the heck do you determine a uavs autonomy anyway?," *Proceedings of the Performance Metrics for Intelligent Systems Workshop, Gaithersburg, Maryland*, 2002.

[3] IBM, "An architectural blueprint for autonomic computing," *IBM Whitepaper*, June 2006.

[4] Autonomic Research Group, University of Greenwich *http://cms1.gre.ac.uk/research/autonomics/tech.html. Latest Access: December 20th, 2012*.

[5] ISO/IEC 9126, "Software engineering Product quality (ISO/IEC 9126)," tech. rep., International Organization for Standardization/International Electrotechnical Commission, 1998.

[6] ISO/IEC 14598, " Information technology – Software product evaluation (ISO/IEC 14598)," tech. rep., International Organization for Standardization/International Electrotechnical Commission, 1999.

[7] H. Shuaib and R. Anthony, "Towards Certifiable Autonomic Computing Systems Part I: A Consistent and Scalable System Design," *International Journal On Advances in Intelligent Systems*, vol. 5, December 2012.

[8] R. W. Proud, J. J. Hart, and R. B. Mrozinski, "Methods for determining the level of autonomy to design into a human spaceflight vehicle: A function specific approach," *PerMIS 03. Proc. Performance Metrics for Intelligent Systems, NIST Special Publication 1014*, September 2003.

[9] M. C. Huebscher and J. A. McCann, "A survey of autonomic computing degrees, models, and applications," *ACM Computing Surveys*, vol. 40(3), August 2008.

[10] IBM, "An architectural blueprint for autonomic computing," *IBM Whitepaper*, 2004.

[11] H.-M. Huang, E. Messina, R. Wade, R. English, B. Novak, and J. Albus, "Autonomy measures for robots," *Proceedings of IMECE: International Mechanical Engineering Congress (IMECE2004-61812)*, November 2004.

[12] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICSPART A: SYSTEMS AND HUMANS*, vol. 30, pp. 286–297, MAY 2000.

[13] W. Truszkowski, L. Hallock, C. Rouff, J. Karlin, J. Rash, M. G.Hinchey, and R. Sterritt, *Autonomous and Autonomic Systems*. Springer, 2009.

[14] B. Moore, E. Ellesson, LongBoard-Inc., J. Strassner, A. Westerinen, and Cisco-Systems, "Policy core information model – version 1 specification (RFC 3060)," February 2001.

[15] B. Moore and IBM, "Policy core information model (PCIM) extensions (RFC 3460)," January 2003.

[16] K. Zeilenga and OpenLDAP Foundation, "Lightweight directory access protocol (LDAP): Technical specification road map (RFC 4510)," June 2006.

[17] M. Wahl, A. Coulbeck, T. Howes, S. Kille, Critical Angle Inc., Netscape Communications Corp., and Isode Limited, "Lightweight directory access protocol (v3): Attribute syntax definitions (RFC 2252)," December 1997.

[18] pseudonym67 *http://www.codeproject.com/KB/recipes/pathfinderBypseudonym67.aspx. Latest Access: December 20th, 2012*.

Educational Video Game Design Based on Educational Playability: A Comprehensive and Integrated Literature Review

Amer Ibrahim, Francisco Luis Gutiérrez Vela, Patricia Paderewski Rodríguez, José Luís González Sánchez, and Natalia Padilla Zea

Dept. of Software Engineering
University of Granada
Granada, Spain
{ameribrahim, fgutierr, patricia, joseluisgs, npadilla}@ugr.es

***Abstract*—Design techniques can have an important effect on how video games teach and players learn. The ability to harness these techniques in the design of educational video game can impact the motivation and engagement of playing and learning by creating more options for players to connect with game content as well as to other players. This article focuses on the design phase of the game development process and highlights the role of some techniques that can be used to design a successful educational video game (guidelines and design patterns). These techniques provide information on good practice and form a basis for evaluating the educational video game quality, acting as useful tools for developers to enhance video game playability. To this end, we have presented a set of guidelines and design patterns in order to provide an acceptable level of playability and, in this way, a better player experiences and learning achievement.**

Keywords- Playability; Player Experience PX; Educational Video Game EVG; Guidelines; Design Pattern.

I. INTRODUCTION

There has recently been a great deal of interest in using and designing games for education, integrating education and game design in order to grabbing the players' attention and maintaining attention over long period of time. Accordingly, we must understand the structure of games and use this in teaching, and meeting the needs and player requirements due to the different profiles of the players on the market. Designing successful Educational Video Games (EVG) requires ensuring their fun and educational aspects and the different design perspectives (artistic, ludic, social, etc), also, requires design techniques that ensures the Player Experience (PX) improvement and provides an appropriate level of fantasy, immersion, learning and challenge to engage players. A poorly designed game or a bad choice of game elements (story, challenges, puzzles, etc.) means that the player spends more time ascertaining how to play than in achieving the objectives of the content provided [3].

In this work, we aim to use guidelines and design patterns techniques, in addition to take into account the playability concept in EVG due to its capability to ensure fun and entertainment of the game, as well as the effect that playability has on PX, and its role to improve the quality of game design [33][34]. Using playability and video game guidelines are useful in order to ensure player satisfaction, and to ensure the development of an effective EVG both from educational and fun standpoints. Also, using design patterns is appropriate as an effective model to support the design and analysis of EVGs so that the video game experience and the efficiency of the learning process may be improved. Our proposal gathers together what is suitable and useful for EVG design (rules, gameplay, PX, etc.) from a set of works that relate to different video game genres, interactive systems, hypermedia systems and multimedia systems.

Following this introduction, in Section 2, we discuss some studies of EVGs design; present the different design technique of EVGs. In Section 3, we present a definition of Educational Playability, in addition to present an analysis of the relationship between playability in EVG and the presented design techniques (guidelines, design patterns), and define the importance of guidelines and design pattern in the design of EVGs. In Section 4, we present our set of playability guidelines and playability design patterns. Finally, we present our conclusions in Section 5, followed by our references.

II. EDUCATIONAL VIDEO GAME DESIGN: RELATED WORK

Many researchers have discussed the use of video games as useful educational tools to improve learning and performance, and they have attempted to provide a set of techniques for EVG design. In this section, we propose some researches that have been realized to provide norms and rules during the design process, which can be useful as a base to introduce a suitable and a high quality set of guidelines and patterns for the design of EVGs.

Malone and Lepper [68] provided that endogenous integration of educational content into games enhanced learner motivation. In a well-developed educational game experience, "some integral relationship [exists] between the instructional content and motivational embellishments". Prensky [44] outlined the principles of good game design in his book, Digital Game-Based Learning. These principles are the heart of what makes digital video games so engaging and addictive. Prensky predicts that games will be "much more realistic, experiential, and immersive" and include "more and better storytelling and characters". He has argued that video games can help provide such a context for learning. Prensky highlighted several relevant concepts that characterize video games: rules, goals, outcomes, feedback, competition, story,

challenge, opposition, problem solving, interaction, and representation.

Kasvi [29] listed the seven requirements suggested by Norman [22] for an effective learning environment (feedback, goals, motivation, challenge, satisfaction, engagement and suitability); Kasvi suggested that computer games fulfill all of the requirements and believes that they "satisfy them better than most other learning mediums". Rosenzweig and Vanderdonckt presented the benefits of using guidelines in HCI [32]. They advised that consistency be ensured among products and services in order to provide a better user experience. Guidelines should be more than one person's lightly-considered opinion, but they are not rigid standards that can form the basis of a contract or a lawsuit. Guidelines are not a comprehensive academic theory that has strong predictive value rather they should be prescriptive [59].

Pivec [43] in the book Guidelines for Game-Based Learning aimed to help all pedagogues, teachers and trainers implement their own ideas in the form of an EVG. Pivec and Kearney [42] have presented a model for game based learning, which includes some Game Flow criteria for player enjoyment in games which can be considered guidelines of the game design based on the following points: Concentration, Challenge, Player Skills, Control, Clear Goals, Feedback, Immersion and Social Interaction. Padilla Zea presented a set of guidelines for designing collaborative educational videogames, which facilitate the incorporation of collaborative processes into educational videogames [52]. Moreover, various works have presented evaluation heuristics which offer mechanisms to improve the PX [26][27], and these heuristics can be helpful to build EVG guidelines.

As previously mentioned EVG is a play and learn environment, this dual nature of EVGs makes them difficult to design and implement. Due to this complex nature of EVGs, it is important to take advantage of what is already known about best practices for each EVG components. We have therefore focused on existing guidelines found in similar works such as e-learning and videogames to build our set of guidelines [39][57].

Also Design patterns in video games have been a topic of discussion for some time now. Using patterns in EVGs allows the integration of educational content and video game design ideas, and in this way they help us to balance the game challenges and the learning objectives in EVGs. The idea of applying the design patterns approach to produce game designs was first described by a practitioner within the game industry [8] and an extensive collection has been developed for gameplay design patterns [61]. Rogers [65] provides a complete guide to video game design from ideas to characters, mechanics, and level design. Church proposed tools to help designers understand game design and to maximize the player's feeling of involvement and self [21].

Björk and Holopainen [61] defined design pattern in their book "Patterns in Game Design" as the following: "Game design patterns are semiformal interdependent descriptions of commonly reoccurring parts of the design of a game that concern gameplay". They also presented a large collection of game design patterns that were compiled by analyzing existing games, explaining the template used for the game design patterns that followed, and suggesting means for identifying patterns and applying them to the design of a game. Church introduced formal abstract design tools (FADTs) as a way to achieve a shared design vocabulary [21]. Falstein attempted to find a list of 400 rules that apply to game design by including rules that make a good game [51]. Church and Falstein have proposed the same objectives; to define a way to describe and share game design knowledge. In his book "Art of Game Design: A Book of Lenses", Schell presented a hundred ways to look at game design from a multiplicity of angles [30]. Several other books have also been written about game design [5][56]. Kiili [36] considered that the objective of patterns was to fulfill the need for a common tool to facilitate the interaction between designers and to develop high quality educational games. He defined pattern design in educational games in this way: Educational game design patterns are semiformal interdependent descriptions of commonly reoccurring parts of the design of an educational game that concern and optimize gameplay from an educational perspective focusing on the integration of engagement and learning objectives.

In this section, we have focused on existing guidelines and patterns found in the works presented above (e-learning systems, EVG requirement, and video game design). By analyzing these works we can conclude that game designers to get a 'good' EVG design, they need to focus on playability, internal structures (Game rules, Game mechanics…), and expected player experience (How games evoke emotional-intellectual responses from players...). Accordingly, educational video game design is suffering the lack of theoretical and methodological norms, and there is a misunderstanding of the game features that must be included in games and the teaching methodologies that are compatible with game playability. In this paper, we will propose a set of guidelines and design patterns that take into account the need to promote and maintain the playability of an EVG, to build and improve the PX of the game process and facilitate the game development process, these guidelines and patterns were compiled by analyzing the existing game techniques to achieve a good EVG design.

III. PLAYABILITY, GUIDELINES AND DESIGN PATTERNS: A TOOL TO IMPROVE THE PLAYER EXPERIENCE

EVGs design has been a topic of discussion for some time now. The importance of EVGs arises from the failure of e-learning systems and technology enhanced learning process to engage students and keep them motivated to study [44]. Due to the lake of motivation in e-learning systems many designers are shifting towards using interactive learning based on games, e.g., Aldrich [13] introduced more interrelated concepts covered by educational simulations.

Learning through play is currently an effective and attractive educational strategy. Games are fun, learning is hard, and forcing people to learn in games can ruin the fun. Lots of literature exits on why games should be good tools for education, but very little on how to ensure that they are. During our research we found many EVGs that appeared on

the market before their efficacy had been ensured. As a result numerous games have failed, due to the lack of game design methods to convert the objectives of players, educators and designers into reality [64], as well as the fact that analyzing methods are rarely used discuss the level of playability in EVGs in a structured way. Another problem arises from the fact that an EVG is a combination of fun and learning. Thus, a good EVG development process requires cooperation and synergy between game designers and educators in order to ensure a good player experience [53].

To this end, we have proposed a set of player centered design patterns that support EVG design, i.e., game design that places the player experience first and foremost. Björk and Holopainen [61] have mentioned several points that demonstrate the need for design patterns that do not depend on a designer's experience: Problem-Solving for Game Interaction Design; Inspiration; Creative Design Tool; Communicating with peers: Communicating with other professions.

Based on the above, we present our opinion regarding design patterns and guidelines in video game design. Design patterns and guidelines provide many benefits during the design of video games:

- Allow new perspectives for both design and analysis, and provide a network of relations between different game design concepts.
- Design patterns are formal tools used for solving known problems, i.e., they function as a design toolbox. Patterns allow different levels of abstraction in order to address a specific game design problem, and offer the best way to solve issues related to software development using a proven solution.
- Facilitate the development of highly cohesive modules, which may be used many times in different contexts and applications.
- Describe many design decisions that cannot be recorded through the use of primitive methods.
- Patterns have the ability to increase the opportunities for communication and reduce misunderstandings between educators, designers and players, leading to more efficient communication between them.
- Allow some aspect of the system structure to change independently of other aspects.

A. Playability in Educational Video Game

The majority of the presented works in this paper discuss the important role of PX in the structure of video game. The term PX -based on user experience definitions [23] – refers to "all aspects related to the player that are affected by and interact with the playing environment", These aspects represent pragmatic and hedonic features of the process of interaction such as: sensation, feelings, emotional response, assessment, user satisfaction and the experience obtained during playing time [34]. A video game with good playability provides a player with positive experiences of the aforementioned aspects.

The vocabulary of software usability centers on effectiveness, efficiency, and satisfaction (ISO 9241-11). But these aspects don't necessarily add up to 'good' EVGs due to the special feature of the experiential dimensions such as fun, motivation and emotion "It is difficult to obtain knowledge about what players did when playing the game, and how meeting different game design elements affected their experience of interacting with the game" [6]. EVGs should be learning environments that are adaptive, scalable, robust, reflexive, and feature modularity, automation and variability [38]. Thus, we suggested the use of playability to engage, activate and entertain players during the playing time.

Playability in video games is based on usability, but goes much further, it extends the User Experience characteristics with players' dimension using a broad set of characteristics such as motivation, pleasure, curiosity, emotion, and social influences [34]. Thus, Playability isn't limited to the degree of fun or entertainment experienced when playing a game. To this end, we have defined the Educational Playability [2] (Playability of EVG) as "the set of properties that describe the PX in the gaming environment, which main goal provide fun and learning in playable and learnable context, during the entire playing time".

We have previously presented that educational playability isn't limited to the fun objectives, but also takes into account the educational objectives to reinforce the player skills and improve his/her current experience [2][3]. We have presented nine attributes to characterize the educational playability:

"Satisfaction": The gratification or pleasure derived from playing a complete video game or some aspect of it. Satisfaction is an attribute with a high degree of subjectivity. Player Satisfaction can be considered as a measurement to assess EVG as a successful learning system. It is related to the presented content, the used mechanism, educational elements design, and game environmental.

"Learnability": The player's capacity to understand and master the game system and mechanics (objectives, rules, how to interact with the video game, etc).

Effectiveness: The resources necessary to offer players a new experience (fun and learning) while they achieve the game's various objectives and reach the final goal.

"Immersion": The capacity of the EVG contents to be believable, such that the player becomes directly involved in the virtual game world. At an educational level, this property is used to measure the ability of an EVG to present the educational aspects implicitly.

"Motivation": The set of game characteristics that prompt a player to perform specific actions and continue undertaking them until they are completed. At an educational level, motivation to play indirectly produces positive motivation to learn.

"Emotion": This refers to the player's involuntary impulses in response to the EVG stimuli that induce feelings or a chain reaction of automatic behaviors. The educational content in EVGs may provoke rejection by the player, which reduces the motivation for the player to explore the game and thus achieve the educational goals.

"Socialization": The set of game attributes, elements and resources that promote the social dimension of the game experience in a group scenario. From an educational perspective, socialization is the ability to support students learning from one another.
"Supportability": we define this as the ability of EVGs to engage and teach players correctly, and encourage them to continue learning and achieve the learning objectives causing playability as motivational element.
"Educability": We define this attribute as: the educational characteristics of video games that support the user´s ability to be aware, understand and master learning goals.

B. Playability and Design Patterns in Educational Video Game

The integration of playing and learning is the main objective of an EVG. In this article, we highlight the properties of player experience which result in a successful game. Player experience is related to all aspects of interaction between the video game content and the player. As previously mentioned, the goal is to ensure an optimal player experience while blending educational objectives with fun challenges. To achieve this, we suggest a set of design patterns based on the playability attributes as a tool to reduce the complexity of EVG design, as well as to help the player improve his/her experience during playing time. A feature of game design patterns is that adding new patterns does not restrict or specialize the nature of the game, but rather expands it. This is because a pattern describes a particular aspect of playability and its effects on player experience.

We thus present a new taxonomy of design patterns, which were compiled by analyzing the existing exiting game design patterns and the current problems that face EVG designers. Each pattern in our proposed set describes a part of the possible interaction between a player and the game. These patterns are related and when used together they are able to improve player experience and effectively resolve EVG problems. Design patterns provide video game designers with the opportunity to play a powerful role in constructing and improving game playability. Design can be reactionary, responding only to current conditions, or it can be visionary, by presenting solutions to problems yet undefined.

The new set of patterns has been created so that the following points are considered: appeals to both cognition and emotion; improves upon the player's previous experience; fosters creativity and collaboration between designers in order to produce the best player experience possible; presents a game structure that is able to bridge the gap between the required experience and the player experience; facilitates the evaluation of the experience and the effectiveness of the game. Using design patterns should increase designers' experience, helping them to ascertain what is meaningful to the end user and how to present it in the best possible way.

C. Playability and Guidelines in Educational Video Game

Playability and guidelines are closely related, where playability can play the role of guidelines during the game design. Playability is a property that should have high levels in order to keep the game fun and to maintain the player motivated throughout the game. Playability is a qualitative property that can be used during both design and evaluation phase; in the design phase, playability can be considered as a set of guidelines regarding to how to implement the necessary elements to give birth to a desired sort of game play or social entertainment. Also, playability is a tool to measure whether player requirements have been achieved, and to determine the playable-learnable aspects of a game.

At the same time, Guidelines are a set of rules used during the design stages to increase the quality of the generated games, guidelines provide information on good practice and can be used as a basis for measuring the quality of an EVG. Guidelines facilitate decision making by designers and game developers when creating or using different elements in the game. The use of guidelines also ensures the design success and further development of the game, to a certain degree.

The complexity of videogame development requires certain specific guidelines in order to be educational without loss of playability. Thus, guidelines play a pivotal role in achieving the goals of a game and that they are useful in creating a highly playable game design. We can say that guidelines are, specially, to ensure high levels of playability in the game. Accordingly, we present a set of guidelines that have been obtained by analyzing the characteristics that a playable game should have. Based on the above, we define the guidelines in EVGs as the following: "a collection of principles, conventions or directives which describe the educational video game characteristics and are used by developers in all stages of game creation".

D. Guidelines and Design Pattern Importance

Having identified that games exhibit some unique playability problems, e.g., the lack of a player role in the game design or the inability to provide the player with the proposed content, we considered that there was a clear need to develop a collection of design patterns and guidelines that specifically addresses playability and player experience in video games. In this work, we emphasize the properties of player experience that are necessary to obtain a successful video game, where this experience is related to all aspects of interaction between the video game and the player. Accordingly, the solution should ensure optimal player experience and be able to blend the educational objectives with playful challenges, present the game objectives, and provide interesting choices, immediate support and assistance, and an attractive learning environment.

We consider that the design of EVGs is ideal for developing design patterns and guidelines because it is able to:

- Support an innovative approach that effectively integrates and balances fun challenges with educational objectives.
- Support the playability attributes and build an optimal experience by including a game design that motivates players.

- Patterns provide a common vocabulary between game developers and designers in this rapidly expanding field.
- Guide the test and evaluation of the game experience and the efficiency of the learning process.

IV. EDUCATIONAL VIDEO GAME DESIGN

The purpose of the design phase is to give the game designer the ability to design educational video game based on the generated requirements in the analysis phase, understand the lack points and problems of players, fix the problems and do walkthroughs of design concepts. The designer can brainstorm the game elements such as motivation and concept; narrative context and story; the goals and rules of the game, including game mechanics; and the player interface, including the feedback and modes of interactivity.

Achieving the players' needs and requirements, and the educators' objectives and requirements at design level involves the use of design pattern and guidelines techniques. In this work, we have presented a set of guidelines, which are sound and unambiguous and then they could be presented to designers without the rationale behind them. Also, we have used design patterns in educational video game as a useful way to explore suggestions for good design and to complete the rule of guidelines.

A. Educational Video Game Guidelines

In this paper, we aim to provide an integrated and comprehensive set of guidelines appropriate for EVGs to be efficient, fun and successful. The design of these guidelines is based on the playability attributes, and takes into account the need to promote and maintain playability, as well as to improve the PX during playing time. Also, we have classified the suggested guidelines into groups related to all EVG aspects. This classification aims to clarify the designers and game developers expectations, and identify how players will be evaluated for each objective more clearly, to show more knowledge of the game content as well as to improve learning content structure, consolidate all resources, activities related to the presented objective. We have classified our guidelines in groups, which have been proposed to strengthen our proposal by achieving the educational playability attributes during the design phase. We have presented the guidelines groups as follow:

1) Game Goals

This defines how the playful and educational contents can be presented in an interactive way, which is both easily understood and attractive to the player. Games should provide enticing long-term goals [25]; in addition a short-term goal will be needed to achieve the overall game goal [10]. Game goals should guide player through the game and can be presented implicitly or explicitly, goals need to be presented early and clearly stated, and should be personally meaningful, obvious, and easily generated [14][69][68]. EVG should offer a particular strength and sustain of motivating users, "if computer games are intrinsically motivational, then they can be exploited to make learning more motivating and learning will happen almost without the individual realizing it" [55]. Thus, appropriate game goals keep players immersed, interested and motivated in playing and take into accounts the prior skills and experience of the player. Also, game goals describe how to provide the player with the possibility to understand and master the game in a systematic, creative manner. These goals can be divided into two categories as the following:

a) Playful Goal: presents the general objectives of the EVG in a simple and enjoyable way that can help to captivate the players.

- Must facilitate the learning process.
- Provide players with clear knowledge about the educational content in order to make the goal easier to achieve.
- Game content must be appropriate for the predetermined learning objectives and players.
- Game allows players to be involved in challenging tasks, not trivial activities.
- Game outcome should be unexpected in order to increase the player´s curiosity.
- "A good game should be easy to learn and difficult to master" (Bushnell's Law).
- Main goal should focus on reinforcing the player's skills and improve his or her prior experience.

b) Educational Goals: describe what the educational component in EVG should be and how it should work.

- Provide analytical and critical thinking.
- Make a systematic introduction of educational content.
- Assessment and recognition of prior learning.
- The goals are divided into several sub-goals to scaffold learning: generally the sub-goals are gradually presented to lead learners to the learning objective.
- Educational content should be: Valid and Reliable, Credible, Accurate, Relevant, Balanced and free of bias.

Based on the definition of the playability attributes presented above, this set of guidelines affects on such playability attributes as: Satisfaction, Learnability, Motivation and Educability. These guidelines facilitate the learning and playing process, give players the ability to master the game, overcome the different challenges in order to keep a player immersed and motivated to play.

2) Balanceability

The balance between fun and education is a very important factor in the success of an EVG, and has a great influence during the playing time. "A key problem in the development of educational games is balancing how much of the game is a game and how much of the game is learning" [37]. Law [23] presents some problems of current EVG as a poor balance between playing and learning activities or between challenge and ability. The imbalance leads to a separation of learning from playing, which leads to an EVG failure.

- Game must include a fun factor to motivate the student to achieve the learning outcomes [42].
- Educational elements should be clear but not dominant.
- All contents should be compliant in terms of goal visualizing and achieving.
- All game steps should contain both EVG components (fun and education).
- Keep consistency during all EVG steps by ensuring efficiency in game component visualization.
- Include different ways to present EVG components.
- Balance must reflect as much as possible the player´s state (emotional, psychological, etc).

The balance between EVGs contents will make a game more active and attractive, which will ensure the player pleasure, the efficiency of the game structure and thus players can reach the game goal easily. It thus affects such playability attributes as: Satisfaction, Immersion, Effectiveness and Supportive playability.

Quest Atlantis [60] presents the balance of education, entertainment feature to support academic learning, individual development, and social transformation. In this manner, it integrates principles underlying the development of entertaining games into the design of a learning environment. Also, it entails a rich metagame context through which children perceive their participation as meaningful and engaging. This design includes a lot of the presented guidelines in this group.

3) Game Challenge

Challenges are the part of video game, which keeps players motivated to seek for knowledge in order to provide a solution and continue with the game. Challenges increase the game dynamic by presenting different levels and types of challenges, challenges should be introduced in a way that give players the opportunity to study their behavior, as well as to provide an appropriate challenges to players skills, which can be seen in the second part of Bushnell theorem that addresses the idea of providing players with challenge that scales to the abilities of the player. Games motivate when they challenge players and, at the same time, maintain the "illusion of winnability" [15][54]. Also, increasing challenge keep players engaged, reducing the potential for students to become distracted, diminishing educational engagement [17].

- The game should have different level settings to challenge all types of players, (novice or expert players).
- The challenge should be produced from the diversity of the game's tasks and from the difficulty of these tasks.
- The game has to scaffold players' skills and students' knowledge. If the game is too difficult it is frustrating, if it is too easy, it is boring [42].
- Challenges and the level of difficulty should be matched to player experience in single-player games or multiplayer ones.
- Challenges should be balanced the playful and educational aspect based on the player progress.

This group aims to challenge players' creativity, and their skills. Game challenges have always been a great way to give players' brain a good work out. Challenges and mysteries are video gaming strategies to engage players and could help them enhance cognitive skills, preparing them for weightier game challenges. It thus affects such playability attributes as: Satisfaction, Motivation, Emotion, Supportability and Educability.

Mavis Beacon [50] presents different tests and level of typing speed, each level has different challenges; Mavis Beacon will monitor players' progress as well as to introduce new challenges to help players continue improving their keyboarding skills.

4) Feedback

Players must be informed as to what they are achieving at the educational and playful levels during the game. Feedback needs to be frequent, clear, constructive and encouraging. Feedback also provides an opportunity to give explanatory information, metacognitive prompts, and clues to correct responses [68]. An appropriate feedback should be presented based on the player decision and performance, and should help players actually to learn and to reach educational and playful goals [20]. Players need to know how skills translate into strategies for playing the game [35]. This takes place through the game cycle and system feedback occurs by:

- Feedback should be presented to the player after a number of failed attempts, and to help him to understand why he has failed.
- The games should be able to stimulate the player to know more about the mechanism of the system by giving clear feedback.
- Provide different type of feedback audio/visual/visceral (music, sound effects, controller vibration, etc).
- Feedback should be immediate with the aim of achieving game goals.
- Feedback must guide the player through the environment, emphasizing key points and offering assistance along the way.
- Feedback should allow the player to monitor the mastery of skills or information.
- A successful game must first familiarize the player with the complete educational task in order to begin the learning process.
- Provide feedback message about all player actions, situations (level complete, game finished, etc.) and status (score list, winner, loser, etc.).

Achieving the presented guidelines in this set will increase player enjoyment and will assist in overcoming all challenges and improving the game´s objective assimilation. These guidelines make the game attractive to a player, keeping him or her interested and motivated to develop skills to overcome the game steps and compete with other players. It thus affects such playability attributes as: Immersion, Satisfaction, Satisfaction and Educability.

Lure of the Labyrinth [40] lacks a good and appropriate design of player feedback, when players make a mistake in a game, it just gets corrected for them without an explanation

in hopes to help players, it did not tell them why so even though it corrected them. It gets way too frustrating.

5) Interactivity

It describes how to make the players feel that they are part of a creative and dynamic community, and how to create more powerful interactive experiences and engagement by streamlining players' interaction and motivation during the game time. Swartout and van Lent [72] deemed that the best games are "highly interactive, deliberately generating tension between the degree of control the story imposes and the player's freedom of interaction". Thus, the playing experience can be boring and unchallenging in games with complete freedom of interaction, when the game progress requires too much control; the player becomes a passive observer rather than an active participant [47]. De Freitas [62] proposed that the interactivity of EVG involves applying changes on the game structure and the learning process, how games are designed, developed and used in practice upon the processes of learning and how learning activities are structured in practice. Thus, we propose the following guidelines:

- Game should have clear and simple instructions and rules.
- Game should respond in different ways to correct and incorrect actions (using sound, images, etc.).
- Permit easy reversal of actions. If a player makes a silly mistake, allow the player to reverse the action, unless it would affect the game balance adversely.
- Introduce elements of positive surprise or special events in strategic locations.
- Introduce enjoyable activities that aren't passive.
- Educational elements should be related to the playful ones in the same video game frame.

This group aims to produce interactive experience that motivate and actively engage players in the game process. Thus, players discover and reinforce their abilities and learn new skills with interactive and fun computer games, which will implicitly affect on the achievement of several playability attributes such as: Satisfaction, Learnability, Motivation, Immersion and Educability.

The ReDistricting Game [70] was designed to educate citizens around the issue of political redistricting. It present a good level of interactivity, it provides many different ways to draw the district lines that meet the basic mission requirements. It provides an active process of discovering the rules of the game, combined with the ability to get feedback on the players' map at any time allows the players to explore and try out strategies, slowly refining them as they learn more about the game.

6) Adaptation

Adaptation is a very important characteristic of EVGs. In adaptive games the level of difficulty increases or decreases depending on a player's performance. When the games are adaptive they support learner preferences for different access pathways and allow the learner to find relevant information while at the same time remaining immersed in the game [16]. Adaptation shouldn't be only as response to players' action during the game time, but should play a crucial role in the success of the learning process by triggering the learning patterns. This means the learning outcome is related to player performance, so if a player finds a game difficult to play, he will leave it and the leaning process will fail [66]. Therefore, we present some guidelines that can be useful to adapt certain features to create a tailored experience for each player based on how they learn and why they play games. In the following we present some adaptation guidelines:

- Game should be easy to modify and adapt (difficulty level, sound level, background music, control keys, etc.).
- Educational content should consist of modules that are flexible enough to be readily utilized during the game time with minimal adaptation efforts.
- Educational content should be able to cater to diverse learning styles and motivation.
- Rationale for using a specific style of educational content should be determined by the needs of the players, game steps and game situation.
- Game contents should be adapted to the individual pace of the player.
- The player should not be overwhelmed by the information the game is provided. Provide ways to hierarchically compartmentalize information.

Introducing this group of guidelines adapts the game content (playful and educational) to the individual pace of a player, relates the educational content to players' needs, this gives players the opportunity to construct a personal profile, and encourages him or her to master the game and complete all steps of the newly customized environment. This will encourage the player to spend the maximum time playing, as well as to compete with other players. It thus affects such playability attributes as: Satisfaction, Motivation, Supportability, Socialization and Immersion.

Mavis Beacon [50] provides many personalized lessons, exercises and tests. Many entertaining typing games, also it presents a detailed progress reports assist in identifying strengths and weaknesses. It provides different level to teaching all letters typing each level teaches some letter typing.

7) Game Control

Providing player with the ability to easily control; generates a sense of belonging in the game environment. Players should feel a sense of control over their actions in the game. In fact, control has been determined to be a deciding variable when motivation has been observed to increase over time for instructional games. A controllable environment allows players to build confidence and self-esteem, as well as to extend their potential and natural abilities, earlier and to further extents. Also, the level of control the players have in their interactions can develop a sense of ownership in the game environment [7][62]. Control is defined by the number of choices available to the learner, the presence of contingency, and a feeling of power given to the learner by allowing them to produce very different outcomes [46].

- The player is not required to learn new control techniques during the game.

- The player should be the one in control. Players want to feel in charge of the game-at least in regard to control of their avatar.
- The control should be intuitive and mapped to platform control.
- Don't throw random uncontrollable events, or tedious or difficult input sequences.
- The control over the program is very crucial: the interaction should determine how the learners observe and infer the rules of the system, which are also the subject matter.

The role of this set is increase the player interactivity and integration with virtual game world. This group allows the player more control over the experience, and engages the competitive mind in a positive way, helps players to develop skills and reduce the margin of error, thus player will find the virtual world enjoyable. It thus affects such playability attributes as: Immersion, Emotion and Learnability.

Storm Tracker [67] is a game that allows player to predict the path of a storm presenting a several options that support a player to domain the storm mechanism, players are actively in control and trial and error comes into play when the final storm prediction is made. Also, learning the storm´s terminology allows players a better understanding and controlling of what is going on during the game and it also helps game play move more smoothly.

8) Ethics

This group defines that the game content should be presented within an acceptable ethical framework which has no negative impact upon the user - How appropriate are the attitudes and beliefs embedded in a game? How appropriate are the implied social attitudes and beliefs, e.g., about violence, gender, race? [4]. EVG must act as a tool to motivate the player's understanding and awareness of the content presented, establishing some basic principles of the real world, such as competition and the ability to complete tasks. Several researches show that player feel more hostile after playing violent video games, especially games that simulate real-life situations. There is also evidence that playing violent games can make people behave more aggressively immediately afterwards. In the same time several researchers have mentioned that the violence in a video game motivates and engages players, and have considered violence as part of our real life, and thus a game will lose something of its realism by hiding the violent scenes [19][71].

- Game should not teach anything that may result in dangerous situations in real-life.
- Avoid content that has an impact on physical, mental or moral development (for example, inhuman and sexist content).
- Avoid messages conveying content of an aggressively nationalistic, ethnocentric, xenophobic, racist or intolerant nature.
- Emphasis on acquiring and maintaining competence during playing time.

Introducing these guidelines and achieving them will develop a player's skills and experience, in particular the social experience. In this way, correct and real information will be presented to the players, which help them to take the responsibility to discover and achieve the game content, and thus the player skills will be developed. It thus affects such playability attributes as: Motivation, Immersion and Educability.

Storm Tracker game does not show the real world dangerous consequences or any other aspect of the storm that has negative effects on the player, i.e., dead bodies, etc. We suggest that a game will be better if it presents dead bodies as result of a storm, but these bodies shouldn´t cause to the players any visual damage.

9) Realism

Another advantage of EVG is that has the ability to simulate real-life situations in a way which can greatly facilitate the learning process. Crawford describes a computer game as "a closed formal system that represents a subset of reality" [15]. This means focusing more on the real life simulation part with rewards, customizations and clear objectives. Tashiro and Dunlap [31] have explored relationships between simulation realism and engagement in learning, they consider the impact of visual realism on learning engagement in educational games. Krcmar [41] says that as the games become more and more realistic, the positive and negative effects on children increase, because "Greater realism leads to greater immersion; greater immersion leads to greater effects". Also, Wood [58] find Players rate realistic video games more favorably than unrealistic ones.

- Make real-world situations and simulations available within the game.
- Relate educational content to real-world simulation.
- Real-world simulation should provide to help game contents to be achieved.
- Make story realistic, presenting real sequence of events throughout the game.

Realism is very associated with the playability attributes such as: Satisfaction, Immersion, Emotion, Motivation and Educability. Realism ensures the diversity of motivations and the accuracy of the information. Also, it motivates the players' understanding and awareness of the presented content, establishing some basic principles of the real world to activate their primer experience.

Storm Tracker game is real world simulator, it is a very good teaching tool about hurricanes without being too graphic or visually overwhelming. Concepts are presented in a way that allows the player to see the real world aspects involved in hurricane storm tracking as well as to present realistic outcomes of these aspects.

10) Game Reward

Any activity that people enjoy doing has some kind of reward. Rewards it is a very important way to encourage players to perform better due to the bonuses and advantages. Currently, most digital games are designed so players must actively complete quests based on a reward system [56]. Moon [56] mentioned the role of reward to induce play, and presented a learning model for video game based on the reward system. Nielsen has presented the reward as an

extrinsic motivation that players get a reward for engaging in an activity, and are not motivated by the activity by itself [63]. The intrinsic rewards arise from the process of learning or playing, and the extrinsic rewards arise from results (grades, points, winning, or approval) [9].Thus, connecting the reward system and player motivation to learning process will help each player to have a more optimal game experience.

- The reward should be appropriate and important in motivating the player.
- The reward should be appropriate to the current educational step or level.
- The reward should be related to the player´s progress in achieving the educational content
- The reward should be commensurate with the capabilities of the player.
- The reward should be presented at an appropriate time to engage the player.
- Rewards can be given several times during one game step if necessary.
- Game should present a variety of rewards.

Presenting this group of guidelines affects on the player participation during the entire playing time due it's important to hold the player attention as well as to make games funny and delighting. The rewards guidelines motivate and encourage players to play and learn. Thus, by using this group we can improve the different playability attributes such as: Motivation, Immersion, Learnability, and Educability.

Math Missions [49], the Amazing Arcade Adventure by Scholastic, players earn money for every correct answer. This money can be spent on buying arcades and they even get to run the arcade. These rewards are presented as a way to motivate the learning without really being related to the learning experience.

11) Structuring

This group is related to how the EVG content is presented in a way which motivates the player, and how to introduce challenges throughout each game level creating an imaginary learning space that is engaging and immersive. The structure of EVG should provide entertaining and interesting content, which should be suitable for players´ primer skills and knowledge. Also, game elements and content should be related and the match between them is a very important, where the aesthetic aspect of the game structure guides the player to achieve the game goals, while the learning design of games should make it possible to use them in a modular way and to organize content and sessions in such a manner as to emphasize this learning aspect (personal learning routes, "*reassembling*" of parts of the games into different training paths, etc.) [43].

- Ensure an aesthetic consistency between playful and educational elements, to guide the player through his or her tasks and objectives.
- Each item has its activity that is easy to distinguish from the activity of other items.
- Ensure that the player can quickly recognize the game environment structure without the need for additional help.
- Design the interface to offer defined tasks. The sequences of actions the player is performing should be arranged into a conceptual group of smaller subtasks. Each task completion should be punctuated with an acknowledgment, so the player knows that his or her task has been completed.
- Utilize game elements that can be easily and quickly understood without requiring any additional help.
- Storyline and the interface elements should be appropriate.
- Each player should be able to easily distinguish his or her items in multi-player environment.

The set of guidelines is related to several playability attributes such as: Effectiveness, Motivation, Emotion and Supportability. A good game structure helps players interact in a simple way, and maintains the enjoyment and curiosity of players. This group provides the ability to control and master the game and successfully overcoming all challenges in order to reach the game objectives.

WolfQuest [73] is an immersive 3D wildlife simulation game, Based on real topographic maps of Yellowstone Park and realistic graphics, which lets players join a wolf pack made up of friends in the multiplayer version or seek to perfect their hunting skills and build their own pack in the single player version. The presented design of this game is related to its story and totally captivates players during the playing time. The presented elements of each frame represent the required function as well as to be easy to understand and master.

12) Player knowledge

This can be defined as how to encourage and motivate players to activate and use their prior knowledge and skills, and intends to generate and improve the players experience by providing them with the new content during playing time. EVGs are collections of skills, knowledge, one advantage of the EVG is its ability to build on or to improve the player's current skills, which arises from the pleasure of mastering a new topic or content being learned, and the curiosity about the subject matter. Thus, a way to build the player knowledge is to engage in the game virtual world during game play [18]. Also, this group emphasizes that the quality of the presented content should be correct and effective. Prensky [45] explained how young people can create their knowledge in practice "the kids who play today's 'complex' video games… learn to think: through experimentation and what real scientists call 'enlightened trial and error,' they learn to understand and manipulate highly complicated systems". Prensky understood that "in order to 'beat' their complex games kids must learn, through complex reasoning, to create strategies for overcoming obstacles and being successful – skills that are immediately generizable". Thus, video games encourage players to achieve mastery to challenge, forcing people to adapt and devolve their thinking and strategy.

- Use the previous experience of players rather than oblige them to learn new knowledge.
- Give the player the opportunity to relate his or her knowledge to real life situations within the game environment.
- Each game level should add something new to a player's knowledge.
- The player should be encouraged to reflect upon the newly acquired information and integrate it into his or her existing body of knowledge.

These guidelines ensure that the player's skills and knowledge is increased and his or her ability to achieve the game goal is improved. These guidelines can increase player confidence during the game, and increase his or her ability to compete with other players. The emphasis on real knowledge in these guidelines will encourage the player to concentrate on achieving the game goals. It thus affects such playability attributes as: Learnability, Educability, Supportability, Emotion, Motivation and Socialization.

A Branches of Power [11] game immerses students in the workings of the three branches of government. The player has to think critically about the decisions he/she is making about what to put in bills and whether they are popular and constitutional or not. Thus, the presented content is definitely not passive learning. During the playing time player has to learn basic ideas and words about what it takes for an issue to move from an idea into a law. Players have to make choices and decisions about what values they take on and what ideas to put into bills, they must think about those same things and how they affect the real world. Thus, players must use their experience about the game content as well as to use the presented content to make their decisions.

B. Design Pattern for Educational Video Game

To achieve learning and entertainment in a gaming environment, we have presented an integrated set of patterns must be suggested to support the player experiences based on the playability attributes in EVGs [1][3].

The proposed patterns are documented in a standard format, as solutions to common design problems. We use patterns as a tool for problem-solving, to support creative and effective design, to build a repository of knowledge and encourage reuse of best practices, and as a way to share designers' experience. By using the patterns described in this section, it should be possible to develop a structure that helps build both entertaining and productive educational video games.

To facilitate the development and use of EVG patterns, we have developed a template based on the proposed elements by Christopher Alexander [12], our template consists of the following main elements: Name, Problem, Description, solution, Playability Elements Affected, and Elements of the Game Affected. Also, we have classified the patterns in relation to "Educational Playability." In this way, we associate the proposed patterns with the situations and the game elements which are closed to the player's experience during the game. We believe that this aspect is most important to develop effective and motivating learning games. In this context, we have classified the proposed patterns within a flexible interactive structure based on playability attributes and related to all aspects of EVGs (Table I), as follows:

1) *Interactive Integration:* describes those patterns that focus on EVGs as a combination of fun and educational elements. It presents the structure of EVG objects, where educational and fun aspects are given more emphasis than in other types of games.

2) *Active Support:* describes those patterns that help and support players to understand what they are doing and learning during the game's progress. This discourages the player from stopping the game, and encourages him or her to think about the decisions, actions or strategies that must be taken in the next step.

3) *Knowledge Realism:* describes those patterns that ensure the quality of the EVG content, by presenting accurate knowledge related to the real world. This gives the user confidence as it enables him or her to check the accuracy of obtained information.

TABLE I. PLAYABILITY DESIGN PATTERNS TAXONOMY

Taxonomy	Pattern Problems	Patterns	EVG Elements
Interactive integration	Create an EVG in which all fun and educational aspects are included, Present the educational content indirectly, Produce a appropriate player interface, Generate a good player experience.	Balanced EVG, Interface Structure, Adaptive Content	Tasks, Disposition, Objective, Challenges.
Active support	Feedback, Keep the player informed about his or her status, Present the necessary information to support the game progress. Incentives to reward players	Related Support, Reward	Feedback, Score, Active Reward.
Knowledge realism	Give players new, correct knowledge, Effect of game reality on player experience.	Knowledge Correctness, Game Reality	Reward, Realism, Challenges, Rules.
Beneficial play	Use game activities to teach, Keep player motivated during playing time and progress in the game.	Incremental Learn, Motivated Play	Reward, Challenges.
Knowledge Growth	Support players to become aware of and to obtain new knowledge. Improve player experience and awareness.	Skills Improvement, Embedded Learning	Challenges, Feedback.
Social awareness	Use social aspect to improve player experience.	Shared Experience	Group challenges, Dependence among members.

4) *Beneficial Play:* describes those patterns that provide players with incentives (reward, fun, pleasure) to encourage them to advance in the game, and consequently, in their knowledge and skill acquisition.

5) *Knowledge Growth:* describes those patterns that focus on the use of EVGs to give players new knowledge and skills, and to improve and develop previous knowledge.

6) *Social Awareness:* describes those patterns that present and use the social features of EVGs to facilitate learning or teaching through social activities and their role in strengthening the player experience.

The previous table (Table I) presents the proposed taxonomy of design patterns, the common problems, the suggested solutions in each group, and EVG elements which are related to the discussed problem [1].

In the following we will present a brief description of the presented design patterns:

"Balanced EVG": describe how to create a game in which all fun and educational aspects are included.

"Interface Structure": present solution to produce an appropriate and attractive player interface.

"Adaptive Content": shows how the EVG actively should provide players with the proposed content.

"Related Support": describe how the EVG helps the player's progress in the virtual game world.

"Reward": describe how the EVG provides players with incentives that encourage them to explore the game world.

"Knowledge Correctness": show how to give player new, correct knowledge during playing time.

"Game Reality": present how the useful aspects of game realism can be used and managed to improve player experience.

"Incremental Learn": focus on the EVGs ability to increase the player's desire to achieve the educational content.

"Motivated Play": presents solution to keep players motivated to play and achieve the EVG contents.

"Skills Improvement": support the player to become aware of new game knowledge and obtain it.

"Embedded Learning": shows how the educational content of an EVG can be presented throughout the game.

"Shared Experience": present how the social characteristics of EVGs can affect the Player Experience.

C. Patterns and Guidelines. The Interrelated Objectives and Role

All the presented design patterns and guideline have built based on educational playability and the optimal player experience, taking the educators view point into account to ensure the educational content integration, and the educational objectives achievement. For example we have proposed a set of guidelines to design a balanced educational video game this guidelines aims to help designers to create active game from a both educational and playful standpoints. In the same time, we have presented a new pattern to treat with the problem of an imbalanced EVG, and thus this pattern provides some ways to avoid this problem during the design phase. As we have mentioned above in Section 4.1 and Section 4.2, this pattern and the related set of guidelines affect on some playability attribute as education, effectiveness, motivation. Accordingly, ensuring a high level of playability during the game design will ensure us a good player-centered design.

In the same way, Player Knowledge guidelines have been presented with the aim of using game contents to build or develop the player experience, which has been supported by several design patterns, Skill Improvement, Knowledge Correctness and Embedded Learning, where Skill Improvement and Embedded Learning introduce solution to player previous skills and knowledge development, while Knowledge Correctness ensure the quality of the player new knowledge. Thus, these three patterns have complementary role to the Player Knowledge guidelines. In other words, these patterns provide solutions to some problems that could face players during the playing times. Achieving this group ensures that the player's skill and knowledge is increased and his or her ability to achieve the game goal improved, and thus they promote the following playability attributes: Supportability, Educative, Effectiveness, and Immersion.

Other example of the integrated role of guidelines and design patterns is Adaptation guidelines that aim to provide the game content complexity and challenges based on the level of knowledge that a player has. This set of guidelines is supported by Adaptive Content design pattern, which provide an active way to provide the different levels of difficulty of the game content based on the player performance, in order to keep the player motivated and encouraged to play. This is related to several playability attributes Satisfaction, Motivation and Immersion, and affects on the Education attributes.

V. PLAYER CENTERED DESIGN PROPOSAL FOR EDUCATIONAL VIDEO GAME

To ensure that the proposed design is player centered, we need a playable prototype, which will be tested with the set of players and educators to get the final educational video game. In educational video game prototypes have three purposes, the first is to define how they video game will work from the player interface perspective; the second how the video game content will integrate from the educators perspective; the third is to test on real players.

To create an educational video game that has the ability to provide what has been presented in the design phase we need an iterative and incremental development process, and thus, any iteration provides the possibility to reveal errors and omissions in the requirements, and to evaluate the proposed design, which help game designers and developers to build gradually an effective and high quality game design. The evaluation during any iteration gives game designers an imagination of the uncompleted objectives based on evaluation of the playability requirements, as well as to inform designers to develop or to change the applied guidelines and the patterns solution to complete the new players´ requirements and suggestions for improvements for the game that result during the evaluation step.

At the fist iterations we suggest the use of low fidelity prototypes that it can be produced quickly and does not

require much development effort, low fidelity prototype helps to start the design process immediately, to do quickly the design change, and designers' decisions are validated with players. However, players could find it difficult to take that leap from the somewhat abstract to the real thing. On the other hand, high fidelity prototype can be used after an acceptable level of designers satisfaction, high fidelity are easy to comprehend by the end user but may require a lot of costly development effort.

As the current market has a different players´ profiles that means the team of players would change their requirements many times with the progression of the game development process, this involves creating a several playable prototypes during the successive iteration in a short time, which should realize the new requirements of players, and thus we suggest the use of the low fidelity to keep the programmers efforts during these first iterations. The result of these iterations creates an acceptable game to players, and thus we can give game programmer an acceptable level of the game requirements to create a playable version as high fidelity prototype.

In this paper, we propose a brief description of a Player-Centered Video Game Development Approach, which includes the proposed design techniques and prototype. Our approach aims to introduce a player from the earliest phases of game development cycle, as well as to use the principles of Playability in EVG throughout the different phases of development in order to achieve a high level of quality in Playability, in the same way as with traditional desktop systems.

A. Analysis

Our approach should start with a game specification that includes the requirements of playability deduced from reference to the facets of playability, analyzing which attribute is affected by which specific video game elements [1].

B. Design

In the EVG design phase we have proposed the use of design patterns that support an EVG design and analysis, which are player centered; we mean game design and analysis that place PX first and foremost. We need design patterns that take into account the need to promote and maintain the playability of EVG at the game process. In addition, we focus on the guidelines for EVG design, which provide information on good practice and form a basis for evaluating the EVG quality, acting as useful tools to enhance videogame playability. Guidelines will be also necessary in order to design appropriate and playable elements according to the context of the game or player profiles.

C. Prototype

In this phase, we emphasize the design needs to prototype and to test with real users. It is therefore important to design and develop playable prototypes, and then will be tested to support the adaptation and refinement of the game with all the players as participants. Thus, we can get the playable elements.

D. Evaluation

Our approach emphasizes the use of playability test during all the development process phases to ensure the quality of playability in the final EVG. It is therefore important to evaluate PX in EVG based on the evaluation of playability. This will enhance the overall game experience by summing all values of PX across all playability attributes, as well as to determine which video game elements have more influence on the final experience throughout the development process. Thus, we will propose a specific heuristics for EVG, which will also help to decide if playability guidelines and playability design patterns are effective, as well as to cover all aspects of PX. Using Facets of Playability also helps to check Playability properties in the different phases.

Table II shows the different phases of our proposal for EVG development based on educational playability (see Figure 1).

TABLE II. PLAYABILITY DEVELOPMENT PHASES

Development Phases	**Main Objective**	**Importance**
Analyze	Understand the player's goals of playing an EVG to determine playability requirements.	Include playability tasks in the project plan that have the educational and playful aspects, Create player profiles, Develop a task analysis, Document player scenarios, Document player performance requirements.
Design	Achieving the proposed design based on the generated playability requirements. (Playability guidelines and playability design patterns).	Achieve the players' needs and requirements, Understanding the lack points and problems of players, and fix the problems and do walkthroughs of design concepts.
Prototype	Design and develop a playable prototype.	Present the designers´ understanding of players' requirements, and the used design techniques.
Evaluation	Evaluating playable EVG by using playability characterization. (Playability evaluation , Playability Heuristics)	EVG to be compelling and engaging, evaluation tests if playability characterization and playability requirements carried out, to ensure the success of the EVG to much extent.

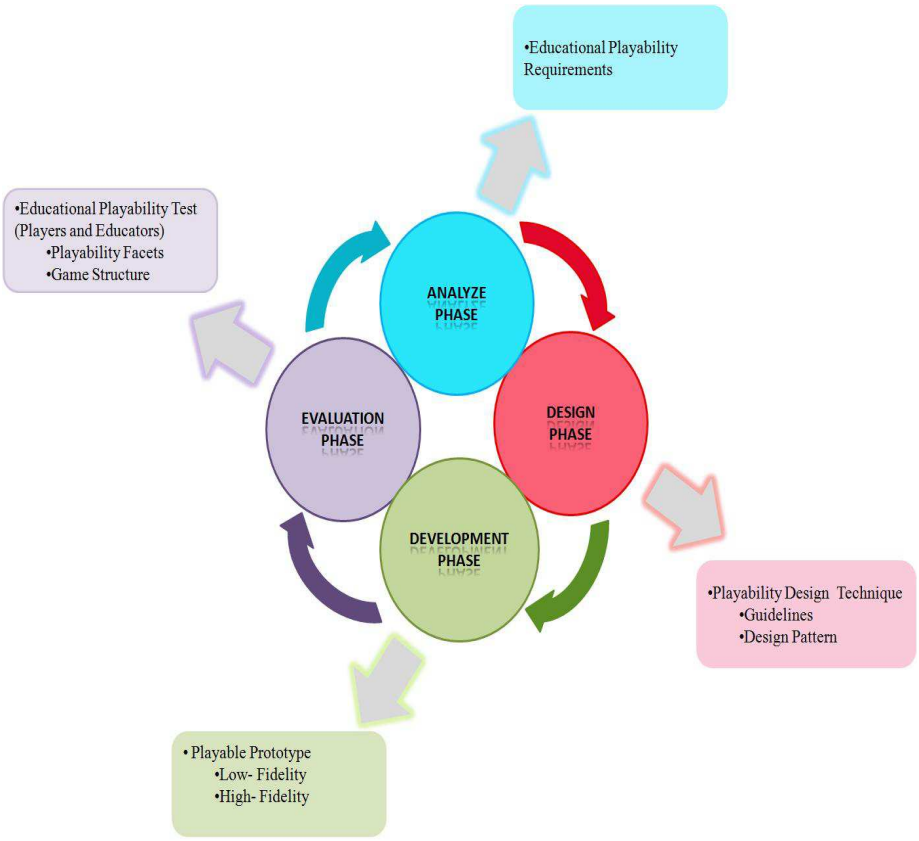


Figure 1. Playability Development Process.

VI. CONCLUSION

Designing and building process of an EVG is far from being a simple task due to the lack of methodologies that provide the necessary constructs and support for the different design tasks. During our research, we have perceived the need for a unified vocabulary and common concepts regarding EVGs and game design. Game design patterns and guidelines are powerful tools that can be used to build a real gaming future, by providing utility, flexibility and scalability of the video game design.

EVGs have the potential to offer real content (playful and educational) in an enjoyable way, which enjoy and build the PX throughout the game progress. The use of certain norms and standards are vital for an EVG to be effective. Guidelines and patterns are presented as certain parameters that make the games apt and useful for their players. Guidelines are a part of game aspects and, in order to be successful, a game must achieve them without losing the playability. Patterns also play a powerful role in constructing and improving the PX by analyzing the exiting game design problems, as well as to be effective to resolve these problems. In this paper, we have proposed some patterns and guidelines; we believe there are many more waiting to be formalized and many more to discover.

Currently, we are working to develop our approach of playability evaluation in order to build a complete catalogue of playability problems, and to develop an extended set of heuristics taking into account the different profiles of evaluators (educators, game designers, etc.). Also, we are working to develop the evaluation approach to be useful to filter the largest number of potential playability issues before making a test with users.

ACKNOWLEDGMENT

This work is financed by the Ministry of Science & Innovation, Spain, as part of VIDECO Project (TIN2011-26928) and the FPU program.

REFERENCES

[1] A. Ibrahim, F.L. Gutiérrez, J.L. González Sánchez, and N. Padilla Zea, “Educational Playability Analyzing Player Experiences in Educational Video Game”, Proc. the Fifth International Conference on Advances in Computer-Human Interactions (ACHI 2012), Jan. 2012a, pp. 326-335.

[2] A. Ibrahim, F.L. Gutiérrez, J.L. González Sánchez, and N. Padilla Zea, “Playability Design Pattern in Educational Video Game”, In González, C.S. (Ed.), Student Usability in Educational Software and Games: Improving Experiences. Hershey, USA: IGI Global, 2012 b.

[3] A. Ibrahim, F.L. Gutiérrez, J.L. González Sánchez, and N. Padilla Zea, “Playability Design Pattern in Educational Video Game”, Proc. 5th European Conference on Games Based Learning (ECGBL 2011), Oct. 2011.

[4] A. Ireland, D. Kaufman, “Computer-Based Games for Learning”, In S. Hirtz, D.G. Harper, and S. Mackenzie

(Eds.), Education for a Digital World: Advice, Guidelines, and Effective Practice from Around the Globe. Vancouver, BC: BCcampus and Commonwealth of Learning, 2008.

[5] A. Rollings, E. Adams, "Andrew Rollings and Ernest Adams on game design (1st ed.)", Indianapolis, Ind.: New Riders, 2003.

[6] A. Tychsen, "Crafting User Experience via Game Metrics Analysis", Proc. the Workshop "Research Goals and Strategies for Studying User Experience and Emotion" at the 5th Nordic Conference on Human-computer interaction: building bridges (NordiCHI), 2008.

[7] A.C. Siang and G.S.V.R.K. Rao, "E-Learning as computer games: Designing immersive and experiential learning" Advances in Multimedia Information Processing (PCM 2004) Pt 2 Proc, vol. 3332, 2004, pp.: 633-640

[8] B. Kreimeier. The Case for Game Design Patterns, 2002. http://www.gamasutra.com/features/20020313/kreimeier_01.htm (accessed on 30/11/12).

[9] B. Magerko, C. Heeter, J. Fitzgerald and B. Medler, "Intelligent adaptation of digital game-based learning", Proc. Conference on Future Play Research Play Share Future Play 08 (p. 200). ACM Press, 2008.

[10] B.W.R. Penuel, S. Pasnik, L. Bates, E. Townsend, L.P. Gallagher, C. Llorente, and N. Hupert, "Preschool Teachers Can Use a Media-Rich Curriculum to Prepare Low-Income Children for School Success: Results of a Randomized Controlled Trial" Education. Education Development Center, Inc. and SRI International, 2009.

[11] Branches of Power http://www.annenbergclassroom.org/page/branches-of-power. (accessed on 30/11/12).

[12] C. Aldrich, Simulations and the future of learning; An Innovative (and Perhaps Revolutionary) Approach to e-Learning. Pfeiffer, San Francisco, 2003.

[13] C. Alexander, H. Davis, J. Martinez, and D. Corner, The Production of Houses. New York, Oxford University Press, 1985.

[14] C. Clanton, "An Interpreted Demonstration of Computer Game Design", Proc. conference summary on Human factors in computing systems (CHI 98), 1998.

[15] C. Crawford, The Art of Computer Game Design. Berkeley, CA: Osborne/ McGraw-Hill, 1984.

[16] C. N. Quinn, "Pragmatic Evaluation: Lessons from Usability", 13th Annual Conference of the Australasian Society or Computers in Learning in Tertiary Education, Adelaide, 1996.

[17] C. Shuler, "Pockets of Potential: Using Mobile Technologies to Promote Children's Learning", New York: The Joan Ganz Cooney Center at Sesame Workshop, 2009.

[18] C.J. Bonk and V.P. Dennen, "Massive multiplayer online gaming: A research framework for military education and training", Washington, DC: U.S. Department of Defense (DUSD/R): Advanced Distributed Learning (ADL) Initiative. 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.153.9995&rep=rep1&type=pdf. (accessed on 30/11/12).

[19] C.P. Barlett and C. Rodeheffer, "Effects of realism on extended violent and nonviolent video game play on aggressive thoughts, feelings, and physiological arousal", Aggressive Behavior, John Wiley & Sons, vol. 35(3), 2009, pp. 213-224.

[20] D. Burgos, "Game-Based Learning and the Role of Feedback: a Case Study", Advanced Technology for Learning, vol. 4(4), 2007.

[21] D. Church, "Formal Abstract Design Tools" Game Developer, vol. 3(8), WIT Press. 1999. http://www1.cs.columbia.edu/~cs4995/files/Doug_Church_FADT.pdf. (accessed on 30/11/12).

[22] D. Norman, Things that make us smarter:Defending Human attributes in the age of the machine. New York, Addison – Wesley, 1993.

[23] E. L.C. Law, M. Kickmeier, D. Albert and A. Holzinger, "Challenges in the Development and Evaluation of Immersive Digital Educational Games", (A. Holzinger, Ed.)Learning, IEEE, vol. 5298, 2008, pp. 19-30.

[24] E. Rosenzweig, "Design Guidelines for Software Products: A Common Look and Feel or a Fantasy?", ACM Interactions, vol. 3(5), 1996.

[25] H. Barwood and N. Falstein, "Better By Design: The 400 Project". Game Developer magazine, vol. 9 (3), March 2002, p. 26.

[26] H. Desurvire, M. Caplan and J.A. Toth, "Using Heuristics to Evaluate the Playability of Games", Proc. the Conference on (CHI 2004), 2004.

[27] H. Korhonen, E.M.I. Koivisto, "Playability Heuristics for Mobile Games", Proc. MobileHCI'06, 2006.

[28] H. MacLeod, J. Haywood, D. Haywood, and F. Littleton, "Choosing and Using a Learning Game in Guidelines for Game-Based Learning". In Pivec, M., Koubek, A. and Dondi, C. (Ed.), Guidelines for Game-Based Learning. Lengerich, Germany: PABST Science Publishing, 2004.

[29] J. Kasvi, Not Just Fun and Games – Internet Games as a Training Medium. Cosiga – Learning with Computerised Simulation Games, 2000.

[30] J. Schell, The Art of Game Design: A Book of Lenses. Morgan Kaufmann Publishers, Burlington, 2008.

[31] J. Tashiro, and D. Dunlap,"The Impact of Realism on Learning Engagement in Educational Games". Proc. of Future Play, 2007.

[32] J. Vanderdonckt, C. Mariage, D. Scapin, C. Leulier, C. Bastien, C. Farenc, P. Palanque, and R. Bastide, "A Framework for Organizing Web Usability Guidelines", Proc. 6th Conference on Human Factors and the Web, Texas, 2000.

[33] J.L. González Sánchez, N. Padilla Zea and Gutiérrez, F.L., Playability: How to Identify the Player Experience in a Video Game. Proc. 12th IFIP TC13 Conference on Human-Computer Interaction (INTERACT2009), 2009, pp. 356-359.

[34] J.L.González Sánchez, Jugabilidad y Videojuegos. Análisis y Diseño de la Experiencia del Jugador en Sistemas Interactivos de Ocio Electrónico (Playability and Video Games. Analysis and Design of User Experience on Interactive and Entertainment Systems). Ed. Académica Española, Lambert Academic Publishing GmbH & Co KG, 2011.

[35] J.P. Gee, "Learning by design: Good Video Games as Learning Machines", E-Learning and Digital Media, vol. 2(1), 2005.

[36] K. Kiili, "Call for learning-game design patterns", In F. Edvardsen and H. Kulle (Eds.), Educational Games: Design, Learning, and Applications, Nova Publishers, 2010, pp. 299-311.

[37] K. Squire, H. Jenkins and the Games-To-Teach Team, "Designing Educational Games: Design Principles From the Games-to-Teach Project", Educational Technology, 2003.

[38] L. Manovich, The language of new media. MA and London: The MIT Press, Cambridge, 2001.

[39] L. Pereira and L. Roque, "Design Guidelines for Learning Games: the Living Forest Game Design Case". Proc. the DIGRA2009 - Breaking New Ground: Innovation in Games, Play, Practice and Theory. West London. 2009.

[40] Lure of the Labyrinth, http://labyrinth.thinkport.org. (accessed on 30/11/12).

[41] M. Krcmar, K. Farrar, and R. McGloin, "The effects of video game realism on attention, retention and aggressive outcomes", Computers in Human Behavior, vol. 27(1), 2010, pp. 432-439.

[42] M. Pivec and P. Kearney, "Designing and Implementing a Game in an Educational Context in Guidelines for Game-Based Learning" In Pivec, M., Koubek, A. and Dondi, C. (Ed.), Guidelines for Game-Based Learning. Lengerich, Germany: PABST Science Publishing, 2008.

[43] M. Pivec, A. Koubek and C. Dondi, Guidelines for Game-Based Learning, Lengerich, Germany: PABST Science Publishing, 2004.

[44] M. Prensky, Digital Game-based Learning. New York, McGraw-Hill, 2001.

[45] M. Prensky, Don't bother me Mom-I'm learning. Minneapolis: Paragon House Publishers. 2006.

[46] M. Westrom, A. Shaban, "Intrinsic Motivation in Microcomputer Games", Journal of Research on Computing in Education, vol. 24(4), 1992, pp. 433-446.

[47] M.J. Dondlinger, "Educational Video Game Design: A Review of the Literature", Journal of Applied Educational Technology, vol. 4 (1), 2007, pp. 21-31.

[48] M.K. Moon, S.G. Jahng, and T.Y. Kim, "A computer-assisted learning model based on the digital game exponential reward system", Turkish Online Journal of Educational Technology, vol. 10(1), 2011, pp. 1-14.

[49] Math Missions http://www.superkids.com/aweb/pages/reviews/math/03/arcade35/merge.shtml. (accessed on 30/11/12).

[50] Mavis Beacon, http://www.mavisbeacon.com. (accessed on 30/11/12).

[51] N. Falstein, "Better by design: The 400 project", Game Developer Magazine, Vol. 9(3), p. 26, 2002.

[52] N. Padilla Zea, J.L. González Sánchez, F. L. Gutiérrez, M. J. Cabrera, and Paderewski, P. "Design of educational multiplayer videogames: A vision from collaborative learning", Advances in Engineering Software, Elsevier Ltd. Vol. 40(12), 2009, pp. 1251-1260.

[53] N. Padilla Zea, Methodology for the Design of Educational Video Games on Architecture for the Analysis of Collaborative Learning. Doctoral dissertation, Granada University, 2011.

[54] N. Whitton, "Encouraging Engagement in Game-Based Learning", International Journal of Game-Based Learning (IJGBL2011), vol. 1(1), 2011, pp.75-84.

[55] N. Whitton, "Motivation and computer game based learning" (R. J. Atkinson, C. McBeath, S. K. A. Soong, and C. Cheers, Eds.) Proc. ASCILITE Singapore, Citeseerx, vol.10, 2007, pp. 1063-1067.

[56] R. Koster, Theory of Fun for Game Design. Paraglyph Press, 2004.

[57] R.E. Mayer and R.C. Clark, E-learning and the Science of Instruction: Proven Guidelines for Consumers and Designers of Multimedia Learning, San Francisco, US: Pfeiffer, 2003.

[58] R.T. Wood, M.D. Griffiths and V. Eatough, Online data collection from video game players: Methodological issues. CyberPsychology and Behavior, vol. 7(5), 2004, pp. 511-518.

[59] Research-Based Web Design & Usability Guidelines. http://www.usability.gov/pdfs/foreword.pdf. (accessed on 30/11/12).

[60] S. Barab, M. Thomas, T. Dodge, R. Carteaux, and H. Tuzun. Making learning fun: Quest Atlantis, a game without guns. Educational Technology Research and Development, vol. 53, 2005, pp. 86-107.

[61] S. Björk and J. Holopainen, Patterns in Game Design. Charles River Media, 2004.

[62] S. De Freitas, "Learning in Immersive worlds: A review of game-based learning", JISC eLearning Innovation. 2006. http://www.jisc.ac.uk/media/documents/programmes/elearninginnovation/gamingreport_v3.pdf. (accessed on 30/11/12).

[63] S. Egenfeldt-Nielsen, "Making Sweet Music: The Educational Use of Computer Games", 2009. http://www.egenfeldt.eu/papers/sweet_music.pdf. (accessed on 30/11/12).

[64] S. Egenfeldt-Nielsen, What Makes a Good Learning Game? Going Beyond Edutainment. 2011. http://elearnmag.acm.org/featured.cfm?aid=1943210. (accessed on 30/11/12).

[65] S. Rogers, Level Up!: The Guide to Great Video Game Design. John Wiley & Sons Publisher, New York, 2010.

[66] S.Thomas, G. Schott and M. Kambouri, "Designing for learning or designing for fun? Setting usability guidelines for mobile educational games", MLEARN 2003 Learning with Mobile Devices, Learning and Skills Development Agency, 2004, pp. 173-181.

[67] Storm Tracker http://gamingandlearning.wikispaces.com/Storm+Tracker. (accessed on 30/11/12).

[68] T. W. Malone and M.R. Lepper, "Making learning fun: A taxonomy of intrinsic motivations for Learning". In R. E. Snow and M. J. Farr (Eds.), Aptitude, learning and instruction: Cognitive and affective process analyses. Hillsdale, NJ: Lawrence Erlbaum Associates, vol. 3, 1987, pp. 223-253.

[69] T. W. Malone, "Toward a theory of intrinsically motivating instruction", Cognitive Science, vol. 4, 1981, pp. 333-369.

[70] The ReDistricting Game, http://www.gamesforchange.org/play/the-redistricting-game/. (accessed on 30/11/12).

[71] W. Bösche, F. Kattner, "Fear of (Serious) Digital Games and Game-Based Learning?: Causes, Consequences and a Possible Countermeasure" International Journal of Game-Based Learning (IJGBL), vol. 1(3), 2011, pp. 1-15.

[72] W. Swartout, and M. Van Lent, "Making a game of system design", Communications of the ACM, vol. 46(7), 2003, 32-39.

[73] WolfQuest. http://www.wolfquest.org/. (accessed on 30/11/12).

Rapid Energy Consumption Pattern Detection with In-Memory Technology

Christian Schwarz,* Felix Leupold,* Tobias Schubotz,* Tim Januschowski,† and Hasso Plattner*
* Hasso Plattner Institute, University of Potsdam
Potsdam, Germany
Email: {firstname.lastname}@hpi.uni-potsdam.de
† SAP Innovation Center
Potsdam, Germany
Email: tim.januschowski@sap.com

***Abstract*—The transformation of today's energy market poses new challenges for both, energy providers and customers alike as the usage of renewable energy sources and energy-awareness increases. Additionally, the energy infrastructure is changing fundamentally. On the one hand, the installation of so called smart meters offers the possibility of more detailed monitoring and fine grained electricity billing. On the other hand, the amount of data produced within the power grid increases dramatically. Utility companies will use such data to increase prediction accuracy and to improve energy production, while consumers will more and more transform to prosumers. Within that environment the necessity of short-term predictions increases to improve the power grids stability. In this article, we respond to some of the challenges that energy consumers and providers face by an implementation of a prototypical recording, monitoring and analysis landscape that uses smart meter data. The challenges that this article tackles include: real-time energy consumption classification; mass energy consumption data classification; and early short-term energy consumption prediction. In extensive experiments on real-world data, we show that such challenges can be handled effectively. We leverage smart meter data via a novel combination of machine-learning algorithms and latest in-memory technology.**

Keywords- energy pattern recognition; smart meters; machine learning; in-memory database; in-memory technology; interquartile range coverage

I. Introduction

The energy sector is currently undergoing transformational changes: Energy providers are facing new challenges and energy consumers are increasingly aware of their consumption and the associated costs.

For energy providers, the rise of renewable energy sources, such as solar or wind energy, drives the evolution from a purely consumption controlled supply network to a production controlled grid [2]. As the ratio of such energy sources increases [3], energy providers must now, more than ever, predict future energy consumption by their consumers earlier on the most up-to-date data in order to match their energy supply with their customers' demand. If energy supply is not sufficient for the predicted demand, the energy provider may decide to buy missing resources from other providers or provide incentives for the customer to change their behavior [4], e.g., via dynamic pricing. In every case, early and accurate energy consumption is essential for the energy provider.

For energy consumers, high environmental awareness [5] as well as economical necessities enforced by rising energy prices [6] drive conscious choices for energy consumption. In the industrial sector, energy expenses can make up to 43% of all operational expenses [7]. With ever increasing energy prices, it is essential for companies to control their spending on energy. Therefore, many companies monitor their energy usage on a more detailed level than most private customers. Companies have successfully reduced their energy consumption, for example, by 58% in the Aluminum industry since 1975 [7]. In the private household sector, a study by the US department for energy showed that simple monitoring energy consumption on in-home displays leads to different consumer behavior [8]. The study showed that 71% of private households changed their energy usage behavior – even if the initial savings only range from 4 to 15% [9]–[11], reported to stagnate at 7.8% on the medium-term [12]. With the increasing number of installed smart meters, private households are expected to monitor their energy consumption on a more frequent basis, leading to an increasing amount of computing power to satisfy the consumers' service needs. Summarizing, it is essential for both energy consumer groups to better understand their own energy consumption behavior.

This article considers pattern detection on energy consumption data. The deployment of pattern detection has the potential to help both, energy providers and energy consumers in their adaptation to the changing energy landscape. Providers can use energy consumption pattern classification to predict upcoming energy usage early on. Pattern detection is particularly useful towards such goals because very often, energy usage is highly regular (e.g., the energy consumption pattern of a manufacturing device). Once the (partial) energy consumption pattern is classified, an already known energy usage pattern of the same class can be used for early energy usage prediction. Energy consumers can use pattern detection to classify their existing energy footprint. Such classification could directly result in savings, for example, by identifying which high-energy consuming patterns occur at high energy price times and subsequently move them to low price times.

Pattern detection on energy consumption data is a big

data challenge by its nature. Additionally, there is a particular focus on real-time data with energy consumption data. The reason for this is the Advanced Metering Infrastructure (AMI) [13] which provides large amounts of real-time data on energy consumption. Processing and storing this data is a challenge in itself – running complex analyses on such real-time data was infeasible with regards to the business value until recently. However, with state-of-the-art infrastructure, such as in-memory technology [14], analytics on large, real-time data is now possible. In-memory technology is therefore a perfect match for implementing pattern detection on energy consumption data.

In short, the contribution of this article is as follows. Using an implementation of in-memory technology [15] as a key component of a monitor, record and analyze environment for energy consumption data, the computational feasibility of energy consumption pattern classification for various use cases is assessed. Extensive experiments show that pattern detection is not only computationally feasible, but also fast and accurate. Classic machine-learning algorithms as well as a purpose-built algorithm (so-called Inter Quartile Range Coverage) for pattern detection are used. Our experiments use real-world energy consumption traces, which were collected for this purpose and published jointly with this article.

This article extends our previous work [1]. Large sections of the original paper were re-written for improved exposition. Additional detail on the data of the experiments is given and the consumption data used in the experiments is published [16]. The main change consists of a more comprehensive experimental evaluation.

The remainder of this article is essentially organized into two parts. In the first part, the foundation for the extensive experimental evaluation that makes up the second part is considered. When considering the foundations for the experiments, Section III presents the used pattern classification methods. Section IV describes the data on which the experiments are conducted. The data collection process by our monitor, record and analyze infrastructure is described and we comment on some characteristics of the data. In the second part consisting of Section V, the pattern recognition algorithms are evaluated in a series of experiments. The conclusion and an outlook on future work are presented in Section VI. We begin with discussing related work and technological foundations next.

II. Background: Related Work and Technological Foundations

In this section, we discuss the background of this article, in particular work related to our research and we describe the technological and infrastructural aspects leading to our decision of using components of in-memory technology as the technological foundation for our work.

A. Related Work

The presentation of work related to this article is structured along the information flow in our experimental set-up, i.e., along the three steps monitoring, recording and analysis. First, energy consumption monitoring and then energy consumption recording infrastructure is discussed, followed by reports on literature on the usage of such information via advanced analytics on such data.

For the considered scenario, smart meters and, more generally speaking, smart grids are fundamental. They are considered to be the continuation of the classical power grid in the information age [4]. In order to avoid different conflicting standards amongst its participant countries, the European Union has instantiated the Smart Meter Coordination Group (SMCG) [17]. In this context, the OPENmeter project has proposed the AMI to be used for the smart grid [18]. Additionally, the Open Metering System has proposed a standard for communication between metering utilities that is independent of the metering devices' manufacturer. Open Metering Systems collaborates with SMCG and also assumes the AMI [17], [19]. Our experimental architecture to monitor and collect energy consumption data in Section IV is similar to the AMI.

Smart meter data will typically be stored in some database. For a number of years, in-memory databases and in particular the so-called in-memory technology, which makes heavy use of column-oriented in-memory databases, have received considerable attention in the literature [1], [14], [20], [21]. In-memory technology can be used to access smart meter data quickly, even though formerly, a column-oriented data layout was perceived not to be well suited for write-intensive workloads originating from a smart grid. To enable the handling of write-intensive workloads as they occur within an AMI, column-oriented in-memory databases use techniques like write-optimized differential buffers and bulk loading [20], [22]. Such techniques have proved that column-oriented in-memory databases are also useful in write-intensive workloads. Apart from storing and providing access to data, in-memory databases have been shown to provide additional benefits: a recent trend in the literature is the usage of in-memory technology for advanced analytical scenarios [4], [23], [24]. The present article extends this line of work.

Once stored in a database, collected smart meter data can be used for various use cases. Optimizing consumer contracts [25] and charging [4] are examples.

Another use case is prediction of energy consumption. Predicting the energy consumption for medium and shorter terms has been done for example with SVMs, e.g., [26], and artificial intelligence approaches such as neural networks [27], [28]. Further related work has focussed on comparing different algorithms for efficient pattern matching over event streams [29]. The general literature on machine learning algorithms is very rich. SVMs are discussed, for example, in [30], [31]. Duan and Keerthi discuss various multi-class implementations of SVMs [32]. Shakhnarovich et al. provide an overview of theory and application of the knn algorithm [33].

This article is set apart from existing approaches by two novel aspects. First, it considers algorithms that were previously not discussed in conjunction with in-memory technology. Second, a new and relevant use-case for in-memory technology is considered: energy consumption pattern detec-

tion. We are not aware of other approaches in the literature that combine machine learning algorithms with in-memory technology for energy consumption pattern detection.

B. Aspects of In-Memory Technology

Utility companies store their data within relational databases in so-called utility systems. The database as the single source of truth within the central system is a logical entry point for implementing new applications that work on real-time energy data like the pattern matching algorithms proposed in this article. In initial experiments, we found that conventional, row-oriented and disk-based database systems could not fulfill the interactive performance needs of our industry-scale sized scenario. This is due to the pure size of the collected data (471 million tuples in the database, including 27 million of the appliance used for the experiments). A disk-based column store like Sybase IQ [34], [35] or HP Vertica [36] would have presented a much better fit based on the analytical style queries needed for realising the scenario of interest of this article. These databases provide the usage of an relational model as presented in Section IV-A2, while reducing the required amount of storage space and hard disk seek times by using columnar storage layout and dictionary encoding. As an additional requirement, we want to run the database queries on real-time data with an ongoing stream of new data inputs into the database. This is the main reason for choosing SAP HANA [15] as the database engine because it provides storage space and combines analytical query optimized columnar layout with the insert performance of row-oriented, in-memory databases [20]. Despite the fact that SAP HANA is an in-memory database, data is stored persistently using database logging techniques, e.g., [37].

All algorithms rely on in-memory technology, in particular the in-memory database engine, for managing the in- and output of data. While all collected smart meter readings are stored within the database, the presented algorithms only need to operate on a pre-aggregated subset of the data. They compare the time series energy consumption data against a global repository of energy patterns. Therefore, the database has to select the corresponding data sets of a certain smart meter, aggregate the data onto a certain granularity level and present the data to the algorithm. Naturally, this step is included in each execution of the algorithms.

III. Pattern Recognition Algorithms

The field of pattern recognition studies the automatic discovery of similarities in data by using machine learning techniques. Pattern matching can be used for regression analysis and classification [30]. Regression analysis fits a function on a set of data points with the goal to extrapolate this data set. The task in classification is to decide whether a data point belongs to a certain class (or not). In this article, focus is on classification.

Supervised learning methods are used in the following. Contrary to un-supervised learning methods, a (mostly) correctly classified training set of data points is given [38]. The goal is to match time series consisting of energy consumption points to energy consumption classes. A classifier gets trained on the training data and is later on used to to classify time series that are not contained in the training set.

More formally, for an input vector $\vec{x}$, a classifier y that correctly classifies $\vec{x}$ into its class $\mathcal{C}$ is build. The existence of K classes is assumed. Supervised learning uses a learning phase, where it is given a set of training vectors X with the corresponding classes. The goal of the learning phase is to construct a (hopefully robust) classifier y that minimizes the classification error on the training set.

We selected the following three implemented pattern matching algorithms for a more detailed presentation: Inter-Quartile Range Coverage, a multi class support vector machine, and a k-nearest neighbor algorithm.

A. Inter-Quartile Range Coverage (IQRC)

The IQRC pattern matching algorithm was specifically implemented for our scenario to classify recorded patterns. Given a set of training vectors X^k with $|X^k| = n$ that belong to the same class $\mathcal{C}_k$, the upper and lower quartile for each component of each $\vec{x} \in X^k$, $\vec{x} \in \mathbb{R}^d$ are calculated. For simplicity of notation, n is assumed to be an even number divisible by 4. The range between the upper and lower quartile is called *inter-quartile range (IQR)*. For each class, d IQRs based on the training vectors are calculated, one for each component. More formally, let us define $\vec{v}_i^k$ as a non-decreasingly ordered sequence containing all n values from the training vectors X^k for component i. The element j is denoted by writing $(\vec{v}_i^k)_j$. So, $(\vec{v}_i^k)_{n/2}$ is the median of the sequence$\vec{v}_i^k$. We define $\mathrm{IQR}(\vec{v}_i^k) = [Q_1(\vec{v}_i^k), Q_3(\vec{v}_i^k)]$, where $Q_1(\vec{v}_i^k) = (\vec{v}_i^k)_{n/4}$ and $Q_3(\vec{v}_i^k) = (\vec{v}_i^k)_{3n/4}$.

IQRC is used to classify data as follows. Given a vector $\vec{x}$, the number of components of $\vec{x}$ that lie in the IQRs for $\mathcal{C}_k$ for $k \in \{1, \ldots, K\}$ is computed. If a previously set threshold for class $\mathcal{C}_k$ is exceeded, i.e., a set number of components of $\vec{x}$ lies within the IQRs of a class $\mathcal{C}_k$, $\vec{x}$ is classified as belonging to $\mathcal{C}_k$. It is possible that IQRC decides that $\vec{x}$ may belong to more than one class. In order to break such ties, the class is chosen, in which the threshold has been exceeded the most.

For classes with a high deviation amongst its members, the IQRs will be larger than for classes with a small deviation. In order to account for this, the weight of a component i lying in the IQR of a class $\mathcal{C}_k$ is set as

$$\frac{1}{1 + Q_3(\vec{v}_i^k) - Q_1(\vec{v}_i^k)} \,. \tag{1}$$

Note that $Q_3(\vec{v}_i^k) - Q_1(\vec{v}_i^k)$ may equal 0 for certain components i. With the weight as defined in (1), we rate those components that lie in smaller IQRs higher than values that lie in a greater IQRs.

The classifier y for IQRC is then formalized as follows. Let $\delta(k)$ denote the threshold of class $\mathcal{C}_k$.

We have

$$y(\vec{x}) = \arg\max_{k \in K} \left(\left(\sum_{i \in I} w(\vec{x}_i, \vec{v}_i^k) \right) - \delta(k) \right)$$

$$\text{where } w(\vec{x}_i, \vec{v}_i^k) = \begin{cases} \frac{1}{1+Q_3(\vec{v}_i^k)-Q_1(\vec{v}_i^k)}, & \text{if } \vec{x}_i \in \text{IQR}(\vec{v}_i^k), \\ 0, & \text{else.} \end{cases}$$

Recall that if more than one class has an IQRC above the threshold, the class for which the threshold is exceeded the most gets chosen. For a relatively high overlap among the classes, it is challenging to identify a threshold that is exceeded by all positive but by none of the negative examples. In the training phase of our algorithm, therefore $\delta(k)$ must be chosen carefully for $k \in K$. We propose a modified hill climbing algorithm [39] in the following. The optimization goal is to maximize the number of true positives, while false positives should be minimized.

Initially, the threshold for each class is determined such that none of the training patterns are classified. Furthermore, the classes are ordered by the size of the corresponding training sets non-increasingly. We start to decrement the first threshold as long as the number of correctly classified training vectors increases. If no further increase occurs, this threshold is then fixed for the class. We continue with the next class to decrement the threshold as long as the number of correctly classified training vectors increases. This step is repeated until all classes were considered. After processing all classes, a new iteration over the classes is started. Contrary to the first iteration, the thresholds of all other products can be assumed to be at a reasonably good value. The iterations over all products are continued until, for one entire iteration over all products, the number of correctly classified vectors does not improve. In our scenario, this typically happens after four iterations.

To place less burden on the choice of the threshold for correct classification, it would also be interesting to consider another variant of IQRC where proximity to the median is weighted additionally. We leave this idea for future work.

B. Multi-Class Support Vector Machine

We also consider classification by Support Vector Machines (SVMs) [31]. SVMs offer binary classification. In our scenario, we aim at classification into K classes, with $K > 2$ typically. The most common approach to extend SVMs for such multi-class classification is the *one-versus-all* approach [32] which we refer to as MCSVM. In the training phase where we have K classes $\mathcal{C}_1, \ldots, \mathcal{C}_n$ and corresponding training vectors, we create K binary SVMs, one for each class. The SVM corresponding to class $\mathcal{C}_i$ is trained with all training vectors from C_i for its first target and with the rest of the training vectors for the other target.

When classifying a pattern that is not contained in the training set, this incoming pattern is passed to each SVM. Ideally, only one SVM detects a positive result. If there is more than one SVM classifying the input as $\mathcal{C}_i$, then the one with the largest result vector is used. If there is no SVM classifying the input as $\mathcal{C}_i$, the one with the smallest negative result vector is chosen. Assuming K classes, this approach needs to test K SVMs.

C. K-Nearest Neighbor

When looking at our energy consumption data, we noticed that the energy consumption patterns have a considerable variance, even if they belong to the same class. Clustering energy consumption patterns into their corresponding classes leads to rather big and possibly even overlapping clusters. Therefore, we also consider classifying an energy consumption pattern by looking for the pattern that is most closely related to the pattern to be classified. The intuition for this is as follows. In a subspace with many energy consumption patterns of class $\mathcal{C}_1$, a pattern of class $\mathcal{C}_2$ varying from the others might still be identified as one that more closely resembles the input and should therefore be chosen. This is what the k-nearest neighbor algorithm does. In the following, we refer to this algorithm as the *knn* algorithm. In our case, it suffices to set k to 1. Given an input vector to be classified, the knn classifier returns the class of the training set element for which the distance is minimal [33]. For simplicity and the fact that our vectors represent continuous variables, we use the Euclidian distance as metric.

An advantage of the knn algorithm is the potential for speed-up by parallelization. Also, there is no computationally expensive learning phase required.

IV. Smart Meter Data

The evaluation of the algorithms described in Section III requires an appropriate data set. While some smart meter data is publicly available, e.g., [40], such data sets are typically too small to be useful in experiments on industry-scale data, which are the focus of this article. Since we also did not want to rely on artificially generated data for our experiments, we decided to record smart metering energy consumption data ourselves. In this section, we describe the experimental set-up for collecting the energy consumption data. We also describe some characteristics of the energy consumption data and the used data model. We decided to make the energy consumption of the used appliance data publicly available [16], in order to, hopefully, facilitate future research in the area of energy pattern classification.

A. Data Collection

Recording real-world energy consumption data presents a challenge in itself [19]. We responded to this challenge by setting up the following experimental environment. We monitored all electric energy consuming devices inside a shared space of our research group and recorded their energy consumption over a period of three months.

1) Monitoring: The electric devices that we monitored represent a subset of the major devices in households: television and home entertainment components, regular illumination, IT components, two fridges and a coffee machine. In our experiment, we measure all devices independently.

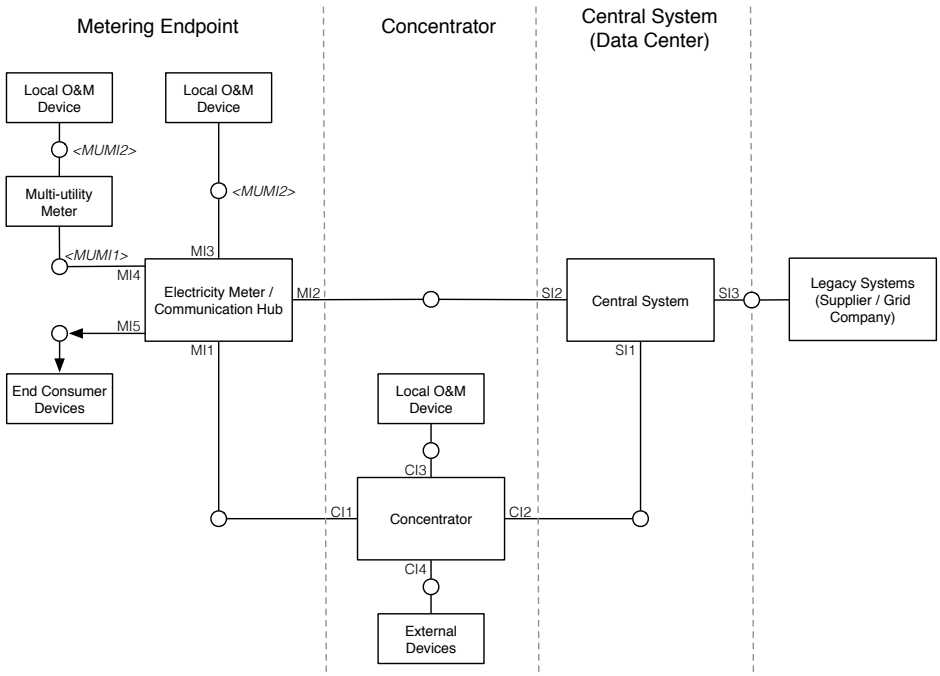


Figure 1. FMC Illustration of Advanced Metering Infrastructure, as defined in [13].

With the costs of self-measuring devices in mind, we decided to use an Emerson Network Power Rack Power Distribution Unit (PDU) with a Liebert MPX control module [41] for monitoring the energy consumption. This PDU is capable of measuring energy data for each connected device independently, e.g., power, voltage, current and rating. The measuring accuracy of the PDU is $\pm 10\%$. Throughout our experiments, we will use the consumption data of the coffee machine for pattern detection purposes. From our collected data, we chose the coffee machine because it presents the toughest challenges for our algorithms. For example, energy consumption varies depending on the coffee being made whereas the energy consumption of a fridge does not vary as much. Both, private households and industry have started to adapt smart meters in order to monitor energy consumption [42]. We expect smart meter to become more prevalent in the future.

2) Recording: For the recording of the energy consumption of the devices, we created an AMI-like multi-level architecture [19], where the energy consumption readings of devices are transmitted via a concentrator component to a central system. In our case, we chose an in-memory database as this central system. This architecture closely resembles the AMI that is expected to be applied for collecting energy consumption data within the power grid of the future. We depict the AMI in Figure 1.

Figure 2 contains a schematic view of our recording infrastructure. The PDU itself is connected to a local area network and data is retrieved using the Simple Network Management Protocol.

The data collector queries the PDU in average once per second to collect the data for each device. Based on these physical measurements, we calculate the power consumption. Finally, we transfer the energy consumption data into the in-memory database. The resulting transmission interval from PDU to database is between 0.5 and 2 seconds depending on the current traffic on our local Ethernet network.

Table I
SCHEMA OF THE TABLES USED FOR PREDICTION.

DEVICE_READINGS	
DEVICE_ID	INTEGER
DATETIME	INTEGER
CONSUMPTION	FLOAT

PATTERN_RECOGNITION	
DATETIME	INTEGER
CONSUMPTION	FLOAT
PRODUCT	FLOAT

We store the collected data in a table called `device_readings`. The occurrence of a pattern is recorded in `pattern_recognition`. The entire database schema is depicted in Table I.

B. Training Set

As mentioned in Section III, supervised machine learning techniques are used for energy pattern detection in our work. Therefore a set of correctly classified energy consumption pattern is needed that can be used to train the algorithms.

The classification challenge for the coffee machine is to detect the type of product that the coffee machine produces, e.g., cappuccino, hot milk or espresso, based on the energy consumption. During the beginning of our data monitoring

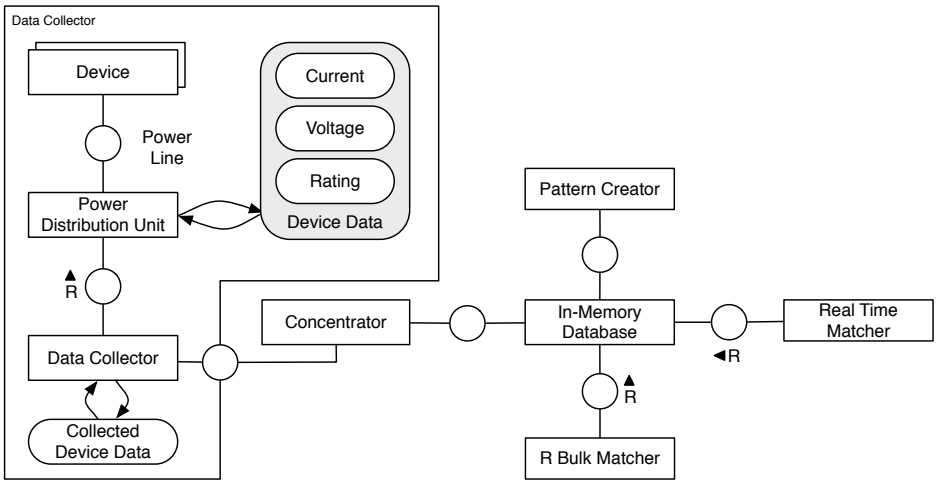


Figure 2. Experimental setup as FMC block diagram.

and recording phase, users of the coffee machine were asked to input the type of coffee product they had selected into a purpose-built application next to the coffee machine. This classification is stored together with the energy consumption data and thus creates a training set for our supervised machine learning algorithms.

Before using this data as a training set for the algorithms, a simple data cleansing algorithm is performed to eliminate energy consumption patterns that have more than three times the standard deviation from the other patterns in their class [43]. Tables II summarizes the training set.

Table II
DETAILS ON THE TRAINING SET.

Product	Number of Occurrences	%
cappuccino	9	4
cleaning	33	15
single espresso	10	4
double espresso	13	6
single coffe	58	26
double coffee	7	3
latte macchiato	89	39
only milk	7	3
total	226	100

Figure 3 shows quartiles of two selected energy patterns, latte macchiato and single coffee. We see the difference between both patterns clearly. However, we note that there is also a large spread of values in each pattern. A number of factors may cause this. For example, measuring inaccuracies by the PDU or coffee-machine inherent reasons such as the coffee water having varying temperatures. It is this large spread among the patterns which makes classification such a challenge.

C. Publicly Available Consumption Data

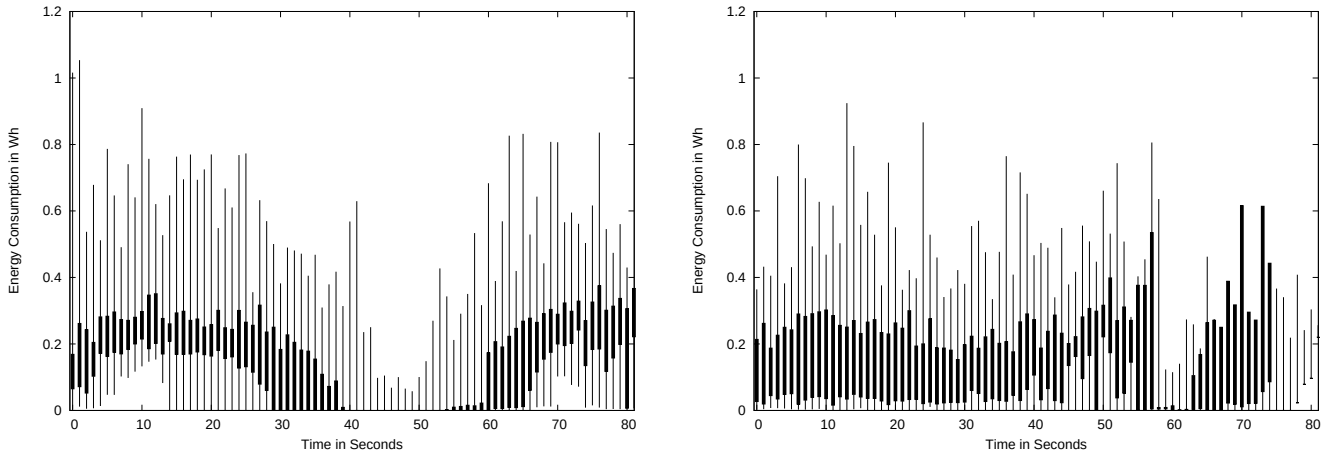
A condensed energy consumption data of the coffee machine used in our experiments is now publicly available [16]. We removed most records that correspond to an idle state of the coffee machine. Table III contains details for the database schema of [16]. Apart from the energy consumption details, the data also contains the classification, whenever it is available. Even though our data comes from a comparably narrow scenario, we think that more general conclusions based on this data are possible. First, the data consists of real, noisy energy consumption. Therefore, it is more appropriate than randomly generated data. Second, the energy consumption traces present a considerable variety of patterns as we noted in Section IV-B. Third, and most important, the sheer size of the data allows for direct conclusion on the computational feasibility of energy pattern detection use cases. The original data set of the coffee machine consists of roughly 27 million tuples, which correspond to one month history of current smart meters for roughly 10,000 households.

Table III
SCHEMA OF PUBLISHED ENERGY DATA.

DEVICE_READINGS		
UNIXTIME	FLOAT	Time in seconds from 01.01.1972
DEVICE_ID	INTEGER	ID of the device at the PDU
POWER_STATE	INTEGER	Power state of the output (2=on)
POWER	INTEGER	Power at the PDU output in W
VOLTAGE	FLOAT	Voltage at the PDU output in V
AMPERAGE	FLOAT	Amperage at the PDU output in A
PATTERN	INTEGER	Identified pattern

V. EVALUATION

In this section, we evaluate the algorithms from Section III on the data presented in Section IV. In our experiments, we evaluate in-memory technology for different energy consumption pattern detection use-cases such as: computational feasibility of on-line pattern classification, accuracy of classification and short-term prediction. Before going into details of the experiments, we give further details on the experimental set-up.



(a) Latte Macchiato. (b) Single Coffee.

Figure 3. Example Quartile Plots for Coffee Products.

A. Preliminaries

For our experiments, we use a HP ProLiant DL580 G7 series server that is equipped with four Intel Nehalem X7560 CPUs and 256 GB main memory. The server runs a 64-bit version of openSUSE 11.2 (kernel 2.6.31.14). We use an instance of SAP's implementation of in-memory technology called HANA [15]. As mentioned in Section II-B, among HANA's features is a column-oriented data layout that is particularly well suited for analytical workloads. Our implementation uses the System R [44] interface of the database development version. The tight integration of R into HANA is a further reason for choosing in-memory technology. Our implementations of knn and MCSVM rely on the rminer package [45].

1) Classification Accuracy Benchmarking: When we benchmark the classification accuracy of our algorithms, we divide our training data set into two parts. One set is used for training the algorithms, while the other set is required for testing the accuracy by comparing the output classification of the algorithms with the actual classification.

We use a cross validation technique to achieve reliable results. The pattern set of each class/product is split into five parts. Each iteration of the algorithm, uses four parts for training and one for testing purposes. The overall performance is calculated by taking the average over all iterations. This technique is called *leave-some-out cross validation* [46].

2) Data Features Used for Classification: In order to classify energy consumption patterns, we use a set of features of each pattern to run our classification algorithms on. Complementing the raw data for electronic power consumption in watt hours, we additionally consider the following features of the gathered data. Let $\vec{x} \in \mathbb{R}^d$ be an energy-consumption pattern.

- Number of peaks: the number of local maxima in $\vec{x}$, i.e., $|\{x_j : x_{j-1} < x_j \wedge x_j > x_{j+1}, 1 < j < d\}|$.
- Greatest Delta: $\max_{i=1,\ldots,d} x_i - \min_{i=1,\ldots,d} x_i$
- Sum: $\sum_{1 \leq i \leq d} x_i$
- Duration: d
- Moving Average: a time series $\vec{a} \in \mathbb{R}^{d-k}$ with $\vec{a}_i = 1/k \sum_{1 \leq j \leq k} x_{i-j/2}$ for appropriate i and suitably chosen k. Within our experiments, $k = 4$ produced the best results, reducing the effect of outliers on the local estimate without over-smoothing the time series.
- Histogram: a sequence of occurrences of distinct energy consumption values ordered non-decreasingly by energy consumption values.

B. Computational Performance of Real-Time Classification

In the first experiment, the computational feasibility of classifying energy consumption patterns is evaluated. Computational speed is important because of the following reasons. Fast response times of our classifiers are mandatory to enable close to real-time matching. The faster the response time is, the earlier the short-term demand can be predicted. This may have direct monetary consequences. Furthermore, only a fast classification allows interaction for which the response limit is 2 seconds [47]. Therefore, queries that are triggered by real-time classification have to be answered within this time interval.

Real-time classification is implemented as a background process that performs classification cycles periodically. When the coffee machine is idle, one cycle takes about three milliseconds. If the coffee machine is not idle, i.e., is currently producing, a cycle still takes less than a second. Table IV shows the cycle times for the different algorithms. It can be seen that knn is the fastest algorithm, on average as well as in the worst case. Matching with MCSVM takes about 30% longer. IQRC needs about twice the time compared to knn. Overall, our experiments clearly show that all algorithms have satisfying performance, as they are well below the critical limit of 2 seconds.

C. Accuracy of Pattern Classification

Our next experiment tests the ability of our algorithms to classify energy consumption patterns after they have been

Table IV
COMPUTING TIMES FOR OUR ALGORITHMS FOR REAL-TIME CLASSIFICATION: AVERAGE AND EXTREME COMPUTING TIMES.

Algorithm	Shortest time in ms	Longest time in ms	Average time in ms
knn	3	806	9
MCSVM	3	1116	10
IQRC	3	1547	13

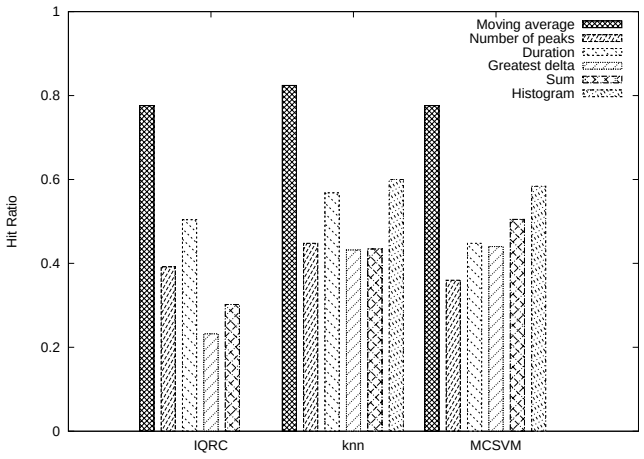


Figure 4. The comparison of three algorithms with different data features.

observed in full. Figure 4 shows the ratios of correctly classified patterns over all patterns for all data features from Section V-A2. We refer to this ratio as the *hit ratio*. We test IQRC, knn, and MCSVM with these features. We remind the reader that we use these features additionally to the raw energy consumption data, i.e., the energy consumption time series themselves.

For IQRC, we were not able to perform the benchmark with the histogram feature, because all patterns of one product result in one histogram. They do not have any deviations or quartiles. The lower boundary for the matching performance with eight different products is 12.5% which would be the accuracy of chance. Recall from Section IV that the PDU, which was used for measuring the consumption data, has an accuracy of $\pm 10\%$ for its measurements [41].

As we can see in Figure 4, the composite feature moving average performs almost equally well across all algorithms and outperforms by far all other features. We think that the smoothing achieved by moving average evens out un-natural deviations caused by measuring inaccuracies and therefore is closer to an idealized pattern.

In general, we observe that the richer the feature is, the more information can be used for classification. This results in higher hit ratios. Comparably simple features perform significantly worse than richer ones. They even seem to disturb the classification. Note that the histogram feature could only be implemented for knn and MCSVM. Though the histogram leads to a slightly better classification than other features, it is still noticeably worse than composite features. The reason is that the measured values hardly differ in size, because the histogram has a lot of different values with low frequencies.

If we compare the algorithms against each other, we notice that knn performs slightly better than the other two. It performs about 5 to 10% better than the other algorithms in all features except for the greatest delta and sum feature. For moving average, the IQRC algorithm has the same hit ratio as MCSVM. Nonetheless, IQRC outperforms MCSVM slightly considering the number of peaks and duration features. MCSVM on the other hand has the strongest results for the greatest delta and sum criteria. The implementation of MCSVM using *one-versus-all* is susceptible to mis-classification if all machines calculate a negative result [48]. Due to the high deviation amongst patterns in our scenario, this case occurs more often. Therefore, the overall accuracy of MCSVM is not as we initially had hoped for. However, due to the high deviation among energy consumption patterns, the hit ratio seems to be overall satisfactory.

D. Computational Performance of Bulk Pattern Recognition

Next, we analyze the computational feasibility with respect to computing times for bulk pattern recognition, where we compare the complete history of the energy consumption data with a set of defined patterns. This scenario is interesting for both the industrial and the private sector, because one could gain an in-depth understanding of the underlying mechanisms of energy consumption behavior. Having classified the energy consumption history allows analyses such as: how much energy was spent on which product/device, or which devices are primarily used during times of highest energy prices. Similar to the experiment in Section V-B, computing times are a critical factor to allow human interaction. However, due to the much larger amount of historical data (versus the live data in Section V-B), computing times will necessarily be significantly higher.

In our experiment, we calculated the average time for one cycle in the algorithm over multiple hours of operation. When we repeat the measurement for the different algorithms, we measure the same time slots on different days. We define one cycle as querying the database for new data plus the time used for matching given there is a pattern detected. Note that we also use energy consumption data that is not contained in the training set for this experiment. Due to the large amount of data involved, the use of an in-memory database is mandatory to allow reasonable computation times. For our experiments, accessing the largest set of data takes a few seconds.

Figure 5 shows the execution times of the MCSVM algorithm, depending on the number of readings in the `device_readings` table for different numbers of used cores (we shall comment on the number of cores further below). We chose this algorithm for our experiment because the MCSVM's computational performance is roughly an average of the other algorithms as the experiment in Section V-B showed. The values in Figure 5 represent the averages of ten measurements with a standard deviation of 11%. One reason for the comparatively large deviation in computing times comes from the fact that during our experiments, energy

consumption data was still being loaded into the database system. We did not stop the loading in order to guarantee more realistic experimental settings.

This experiment shows that the computation time for bulk matching grows linearly in the number of records. Note that both axes have a log-scale. This is particularly pronounced for more than 1000 records. Based on this and the low total computing times, we conclude that bulk matching is computationally feasible for data set sizes that match smart grid use cases as we commented in Section IV-C.

In Table V, we give the results of the bulk pattern classification. We do not give statistics on the accuracy, as we have already commented on this in Section V-C and we cannot measure accuracy on unclassified data. However, we note that the distribution of patterns resembles the training set in Table II.

In this experiment, we laid a special focus on parallelization. Note that the *rminer* package, on which the implementation of our algorithms is based, does not parallelize its computation. However, we parallelized the execution of our algorithms by partitioning the data. Each of the CPUs could then independently work on an equally sized fraction of the total values in the energy consumption data.

We remark that, independent of the degree of parallelization achieved (measured in the number of used cores in Figure 5), the computation times grows linearly with the number of records to be classified. This is expected. Also expected is a decrease of total running time for a fixed number of records with an increasing degree of parallelization. Somewhat unexpected is that this speed-up is comparably small. We explain this as follows. According to Ahmdal's law the speedup is determined by the serial fraction of the algorithm [49]. In our case, this fraction is determined by the initialization of the classifier and the merging of different results for partitions of the `device_readings` table. Merging these results for a total of one million data sets already takes 10 to 20 ms. Although we tried to parallelize this merge, the increase from eight to 32 processes even increases the execution time for less than 200,000 values. The overhead in the merge is not outrun by the smaller number of device readings which each process has to analyze. Nevertheless, 32 cores still outperform eight cores for more than 200,000 readings. With an increasing number of readings, we expect the gain from executing the computing expensive operations in parallel to increase further.

Table V
RESULTING PATTERNS FROM BULK PATTERN RECOGNITION.

Product	Number of Occurences	%
cappucino	82	2
clean	409	7
single espresso	819	15
double espresso	491	9
single coffee	1719	31
double coffee	82	1
latte macchiato	1801	33
only milk	82	1
total	5485	100

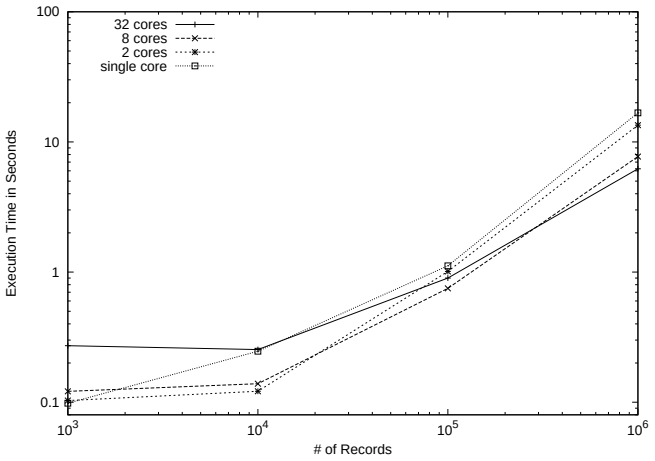


Figure 5. Comparison of execution times for bulk pattern recognition depending on the number of reading for different number of CPU cores.

E. Accuracy of Short-Term Energy Consumption Pattern Detection

In this experiment, we test the accuracy of classifying partial energy consumption patterns, i.e., classifying an energy consumption pattern before it is completed. The motivation for this experiment is as follows. If a pattern is detected, subsequent values of patterns from the same class can be used for predicting the future consumption of a device. The earlier we correctly classify the energy pattern, the more useful this classification becomes as the prediction period becomes longer. However, it is also more difficult to correctly classify patterns, the shorter they have been observed. This is because early classification has to be performed on incomplete energy consumption data and is therefore not as accurate as classification after the complete consumption. Therefore, we need to trade-off classification accuracy with point in time of classification.

Figure 6 shows the accuracy of the knn and MCSVM algorithm depending on the length of the patterns. If we pass a pattern with length n, we cut all training patterns down to that length and apply the algorithms.

We consider a classification rate of 0.5 to be sufficient in order to speak of successful pattern recognition. There are eight possible beverages, a success rate bigger than 50% would be four times better than chance. As we can see, we break the 0.5 accuracy line at approximately 20 seconds. This means that approximately one third of the pattern is sufficient for pattern recognition. If we transfer that finding to industrial manufacturing processes that take multiple hours, the moment of classification is early enough for utility companies, as it provides sufficient headroom for trading, e.g., at the EEX spot market [50].

Since we can classify the energy consumption after twenty seconds, we can predict the succeeding ten to seventy seconds using the information from our trained patterns. We consider this in the next experiment.

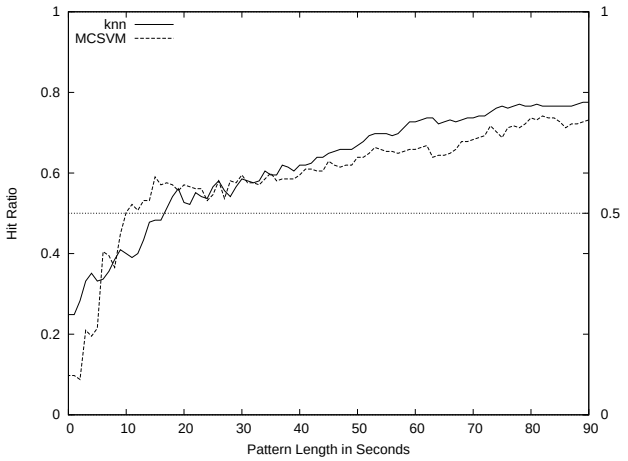


Figure 6. Hit ratio depending on the length of the input vector.

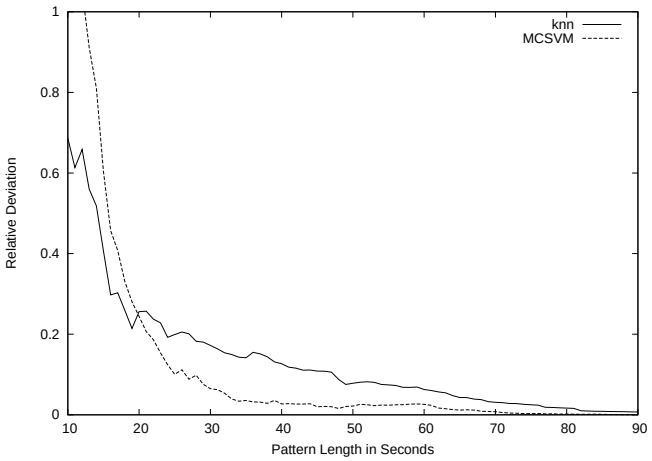


Figure 7. Accuracy of prediction over length of pattern.

F. Accuracy of Short-Term Energy Consumption Prediction

As a final use case of our pattern matching approach for energy consumption, we measure prediction accuracy of energy consumption. Apart from more obvious use cases where an energy provider could try to buy capacities based on her predictions, energy consumers could use predictions for anomaly detection. Whenever we have detected a pattern and its subsequent values differ considerably from the ones predicted, this could either mean that we have a pattern that we have not had in the training set or that something in the device causing the pattern went abnormally. In both cases, issuing a warning seems reasonable. Either the training set needs to be updated or the machine needs to be checked.

We implemented two methods for predicting energy consumption. The first method uses the knn algorithm from Section III-C as follows. We choose the most closely related pattern from our training set to the one having been partially observed. We use this energy consumption pattern from the training set as a prediction of future values for the current pattern. The second method uses the MCSVM method from Section III-B as follows. We start by classifying the observed pattern using our MCSVM method. Next, we identify a training pattern within this class with the least Euclidean distance to our observed pattern. We finally use the values of the training pattern for the prediction.

When we predict the subsequent consumption of a pattern, we have to balance prediction accuracy with time of prediction similar to Section V-E. It is clear that the longer we wait after a pattern has started, the more accurate we can predict the rest of the pattern. However, the longer we wait, the less valuable the prediction becomes. Managing this trade-off depends highly on the setting, e.g., on an economic cost model, and must be decided for the concrete use-case. This is shown in Figure 7. It shows the average prediction accuracy depending on the time the prediction is made. We measure accuracy as the absolute difference of the predicted consumption and the true consumption. This difference is divided by the true consumption.

Our motivation for using this error measure comes from the use case of short term energy forecasting by energy providers. These energy providers could use the total energy consumption of the short term forecast to decide how much energy they would need to provide in the next space of time.

Figure 7 shows that after 20 seconds, i.e., after less than one third of the pattern, we have an average deviation of 25% between prediction and actual consumption. For other prediction use cases, this value seems to be quite acceptable [23]. After 40 seconds, i.e., less than half of the duration of the pattern, the deviation falls even below 20%.

Considering that the consumption values of the coffee machine even under load ranges between 0.1 and 0.8 watt seconds, a predicted value that only differs by .01 watt seconds may lead to a deviation of 10%. Therefore we would have to predict three decimal places correctly to fall below that number. Recall that the accuracy of the PDU is only around $\pm 10\%$ which further complicates predictions. In more advanced scenarios, e.g., for high performance industrial machines, the consumption is higher than for the coffee machine. We expect the precision in measuring energy consumption for industrial use cases to be higher. This may lead to more accurate predictions because the training set may be better.

VI. Conclusion and Future Work

In this article, we presented a case study that suggests that leveraging real-world mass energy consumption data for smart analytics is computationally feasible. An essential component for the success of our case study is the deployment of in-memory technology as implemented in [15]. This in-memory database handles in-coming, live energy consumption data, while, at the same time, allowing analytics on the collected mass data with rapid response times.

As part of the contribution of this article, we make the energy consumption data that we used in our experiments public [16]. We think that the following general conclusions are possible based on the experimental evaluation on our data:

- Our experiments reveal that a number of use cases for energy consumption data analytics can be handled effectively with in-memory technology. Short-term prediction of energy consumption based on short-term classification of energy consumption traces into corresponding patterns is feasible. This opens up many opportunities both for energy providers and energy consumers. Energy providers could use short-term predictions for better trading and classification for better pricing. Energy consumers could use short-term predictions for early warning systems and classification for timing energy consumption better, for example, in order to reduce maximum energy consumption levels.
- While the length of the pattern is rather short in our data, our experiments suggest that after about 20 to 30% of any sufficiently distinct energy consumption trace, it may be recognized and predicted with sufficient accuracy independent of its absolute length. We believe that other energy consumption data sets could be easier to classify into different patterns since the patterns in our data are comparably similar.
- Classification of large real-world sized data sets is computationally feasible with in-memory technology. Such classification allows energy consumers and providers to deeply analyze and understand existing energy consumption data.

Future work may include evaluating the results on other data sets. For example, on energy consumption data from different types of manufacturing machines that produce different and more diverse energy usage footprints. A further possibility for future work would be unsupervised machine learning methods; in particular, benchmarking such unsupervised methods with the presented supervised methods in terms of accuracy and computational speed. Such unsupervised methods would provide useful insights in scenarios where no training data is available.

Finally, our experiments reveal that the speed of in-memory technology-based energy consumption pattern detection is such that machine-human interaction is possible, thus allowing a combination of human insight with machine learning algorithms. Enabling this human-machine interaction for energy consumption classification and prediction would be a most interesting avenue for future work.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their insightful comments that helped improve the article.

REFERENCES

[1] C. Schwarz, F. Leupold, and T. Schubotz, "Short-term energy pattern detection of manufacturing machines with in-memory databases – a case study," in *Proceedings of the Second International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies, ENERGY 2012*, 2012.

[2] P. Evans-Greenwood and A. Mulholland, "Moving the energy industry from demand- to supply-driven," http://www.slideshare.net/peg/moving-the-energy-industry-from-demand-to-supply-driven, 2007.

[3] eurostat, "Renewable energy statistics," http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Renewable_energy_statistics#Electricity, 2011.

[4] M.-P. Schapranow, R. Kühne, A. Zeier, and H. Plattner, "Enabling Real-Time Charging for Smart Grid Infrastructures using In-Memory Databases," in *IEEE 35th Conference on Local Computer Networks (LCN)*. IEEE, 2010, pp. 1040–1045.

[5] eurostat, "Environmental statistics and accounts in Europe," http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-32-10-283/EN/KS-32-10-283-EN.PDF, 2010.

[6] Department of Energy & Climate Change, "Energy price statistics," http://www.decc.gov.uk/en/content/cms/statistics/energy_stats/prices/prices.aspx, 2012.

[7] The NEED Project, "Energy Consumption," *Intermediate Energy Infobook*, p. 46, 2011.

[8] L. A. Butler, "In-Home Display Pilot," *US Department of Energy - Energy Efficiency and Renewable Energy*, pp. 1–9, Jul. 2011.

[9] W. Abrahamse, L. Steg, C. Vlek, and T. Rothengatter, "A review of intervention studies aimed at household energy conservation," *Journal of Environmental Psychology*, vol. 25, no. 3, pp. 273–291, Sep. 2005.

[10] S. Darby, "The Effectiveness of Feedback on Energy Consumption," *Environmental Change Institute, University of Oxford*, pp. 1–24, Apr. 2006.

[11] ——, "Energy feedback in buildings: improving the infrastructure for demand reduction," *Building Research & Information*, vol. 36, no. 5, pp. 499–508, Oct. 2008.

[12] S. S. van Dam, C. A. Bakker, and J. D. M. van Hal, "Home energy monitors: impact over the medium-term," *Building Research & Information*, vol. 38, no. 5, pp. 458–469, Oct. 2010.

[13] The OPENmeter Consortium, "Report on the identification and specification of functional, technical, economical and general requirements of advanced multi-metering infrastructure, including security requirements," *Deliverables*, June 2009.

[14] H. Plattner, "A common database approach for OLTP and OLAP using an in-memory column database," *Proceedings of the 35th SIGMOD International Conference on Management of Data*, Jun. 2009.

[15] SAP AG, "SAP HANA product information," www.sap.com/HANA, 2012.

[16] C. Schwarz, F. Leupold, T. Schubotz, T. Januschowski, and H. Plattner, "Energy consumption data," http://inmemoryeffect.com/energy/data/, 2012.

[17] Arbeitsgemeinschaft Für Sparsame und Umweltfreundlichen Energieverbrauch E.V., "Smart Meter - Intelligente Zähler."

[18] OPENmeter, "Requirements of AMI," Tech. Rep., 2009.

[19] H. Baden and P. Gabriel, "Open metering system specification," Open Metering System Group, Tech. Rep., 2011.

[20] V. Sikka, F. Faerber, W. Lehner, S. K. Cha, T. Peh, and C. Bornhoevd, "Efficient transaction processing in SAP HANA database: the end of a column store myth," in *Proceedings of the 2012 SIGMOD International Conference on Management of Data*, May 2012, pp. 731–742.

[21] H. Plattner and A. Zeier, *In-Memory Data Management*. Springer, 2011.

[22] J. Krueger, M. Grund, C. Tinnefeld, and H. Plattner, "Optimizing write performance for read optimized databases," *Database Systems for Advanced Applications*, 2010.

[23] T. Januschowski, M. Lorenz, C. Schwarz, E. Folkerts, R. Heimburger, A. Akkas, N. Youssef, and D. Simchi-Levi, "Demand forecasting with partial POS data using in-memory technology," in *Proceedings of the 32nd Annual International Symposium on Forecasting*, 2012.

[24] J.-H. Boese, G. Rabinovitch, M. Steinbrecher, M. Magarian, M. Marcon, C. Tosun, and V. Sikka, "Data mining in life sciences using in-memory DBMSs: A case study on SAP's in-memory computing engine," in *Proceedings of Business Intelligence for the Real Time Enterprise*, 2012.

[25] C.-C. Chuang, J. Y. C. Wen, and R.-I. Chang, "Consumer Energy Management System: Contract Optimization using Forecasted Demand," in *ENERGY 2011: The First International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies*, 2011, pp. 58–63.

[26] B.-J. Chen, M.-W. Chang, and C.-J. Lin, "Load forecasting using support vector machines: a study on EUNITE competition 2001," *Power Systems, IEEE Transactions on*, vol. 19, no. 4, 2004.

[27] P. A. Gonzalez and J. M. Zamarreto, "Prediction of hourly energy consumption in buildings based on a feedback artificial neural network," *Energy and Buildings*, vol. 37, pp. 595 – 601, 2005.

[28] S. A. Kalogirou, "Applications of artificial neural-networks for energy systems," *Applied Energy*, vol. 67, no. 1–2, pp. 17–35, 2000.

[29] J. Agrawal, Y. Diao, D. Gyllstrom, and N. Immerman, "Efficient pattern matching over event streams," in *Proceedings of the 2008 SIGMOD International Conference on Management of Data*, 2008, pp. 147–160.

[30] S. Marsland, *Machine Learning: An Algorithmic Perspective*. Chapmann & Hall/CRC, 2009.

[31] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, pp. 273–297, 1995.

[32] K. Duan and S. S. Keerthi, "Which is the best multiclass SVM method? An empirical study," in *Proceedings of the Sixth International Workshop on Multiple Classifier Systems*, 2005, pp. 278–285.

[33] G. Shakhnarovich, T. Darrell, and P. Indyk, Eds., *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. The MIT Press, 2006.

[34] C. D. French, "One size fits all database architectures do not work for dss," 1995.

[35] Sybase, Inc., "SAP Sybase IQ Columnar Database," http://www.sybase.com/products/datawarehousing/sybaseiq, 2012.

[36] Vertica, "Columnar storage and execution," http://www.vertica.com/the-analytics-platform/columnar-storage-execution/, 2012.

[37] J. Lee, K. Kim, and S. K. Cha, "Differential logging: A commutative and associative logging scheme for highly parallel main memory database," 2001, pp. 173–182.

[38] C. M. Bishop, *Pattern Recognition And Machine Learning*. Springer, 2007.

[39] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach (2nd Edition)*, ser. Prentice Hall Series in Artificial Intelligence. Prentice Hall, 2002.

[40] OpenEI, "Energy datasets," http://en.openei.org/datasets/, 2012.

[41] Emerson Electric Co., "User Manual – Network Interface card for the Liebert Rack PDU family of power distribution products," Monitoring For Business-Critical Continuity, Tech. Rep., 2009.

[42] U. E. Information, "How many smart meters are installed in the U.S. and who has them?" http://www.eia.gov/tools/faqs/faq.cfm?id=108&t=3, 2010.

[43] D. Ruan, G. Chen, E. E. Kerre, and G. Wets, Eds., *Intelligent Data Mining: Techniques and Applications*, ser. Studies in Computational Intelligence. Springer, 2005.

[44] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

[45] P. Cortez, "Data Mining with Neural Networks and Support Vector Machines using the R/rminer Tool," in *Advances in Data Mining – Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*. Berlin, Germany: LNAI 6171, Springer, 2010, pp. 572–583.

[46] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, 1995, pp. 1137–1143.

[47] W. O. Galitz, *The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques*. Wiley & Sons, 2007.

[48] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, pp. 415–425, 2002.

[49] G. M. Amdahl, "Validity of the single processor approach to achieving large scale computing capabilities," *Proceedings of the 30th AFIPS 1967 Spring Joint Computer Conference*, pp. 483–485, 1967.

[50] European Energy Exchange AG, "Transparency in energy markets," Tech. Rep., 2011.

Involving All Stakeholders in the Development of TV Applications for Elderly

José Coelho, Carlos Duarte, Tiago Guerreiro, Pedro Feiteira, Daniel Costa, David Costa, Bruno Neves, and Fernando Alves
LaSIGE, Department of Informatics
University of Lisbon
Lisbon, PT
{jcoelho, cad, tjvg, pfeiteira}@di.fc.ul.pt, {thewisher, dcosta, bneves, falves}@lasige.di.fc.ul.pt

Pradipta Biswas and Patrick Langdon
Department of Engineering
University of Cambridge
Cambridge, UK
pb400@cam.ac.uk, pml24@eng.cam.ac.uk

Abstract- **The development of new digital TV systems and the design practices adopted in the development of new TV based applications often isolate elderly and disabled users. By considering them as users with special needs and not taking their problems into account during the design phase of an application, developers are creating new accessibility problems or just keeping bad old habits. In this paper, we describe a novel adaptive accessibility approach on how to develop accessible TV applications, by making use of multimodal interaction techniques and without requiring too much effort from the developers. By putting user-centered design techniques in practice, and supporting the use of multimodal interfaces with several input and output devices, we confront users, developers and manufactures with new interaction and design paradigms. From their evaluation, new techniques are created capable of helping in the development of accessible TV applications, with special interest for a novel way of acquiring and providing knowledge from and to the users with an application called User Initialization Application.**

Keywords–multimodal; adaptation; developers; elderly; user initialization.

I. INTRODUCTION

Ageing is certainly an obstacle to adequate human-computer interaction, mostly because of physical and cognitive impairments. Traditional computational systems only provide keyboard and mouse interaction to users. This makes impossible, for example, for users with severe motor impairments to interact in any manner (at least effectively). Also, as recent developments are responsible for new television (TV) systems and applications, unimodal interaction is still being favored without accessibility concerns, excluding persons whom suffer from an impairment of the sensory channel needed to interact. This situation brings social exclusion and e-exclusion to the Human-Computer Interaction (HCI) world and new TV platforms, as it seriously restricts actions and information access to users with impairments (like the elderly), providing means of interaction exclusively for the so called "normal users". However, multimodality can resolve this issue by offering the possibility of presenting content in many ways (audio, visual, haptic), and in the most suitable way to each user's characteristics. Also, by offering users the possibility to use the inputs more adequate to them (or the context of interaction), in a single or combined manner, multimodal interaction can improve interaction efficiency and, more importantly, accessibility.

Multimodal interfaces aim to provide a more natural and transparent way of interaction with users. They have been able to enhance human-computer interaction (HCI) in many numbers of ways, including: User satisfaction: studies revealed that people favor multiple-action modalities for virtual object manipulation tasks [14]; Oviatt [17] has also shown that about 95% of users prefer multimodal interaction over unimodal interaction; Robustness and Accuracy: "using a number of modes can increase the vocabulary of symbols available to the user" leading to an increased accessibility [15]. Oviatt stated that multiple inputs have a great potential to improve information and systems accessibility, because by complementing each other, they can yield a "highly synergistic blend in which the strengths of each mode are capitalized upon and used to overcome weaknesses in the other" [18]; Efficiency and Reliability: Multimodal interfaces are more efficient than unimodal interfaces, because they can in fact speed up tasks completion by 10% and improve error handling and reliability [16]; Adaptivity: Multimodal interfaces also offer an increase in flexibility and adaptivity in interaction because of the ability to switch among different modes of input, to whichever is more convenient or accessible to a user [15]. However, Vitense [20] illustrates the need of additional research in multimodal interaction, especially involving elderly people. This paper tries to extend this knowledge.

Also, the majority of current approaches to the development of multimodal or adaptive systems, either addresses specific technical problems, or is dedicated to specific modalities. The technical problems dealt with include multimodal fusion [10], presentation planning [10], content selection [12], multimodal disambiguation [18], dialogue structures [3], or input management [9]. Platforms that combine specific modalities are in most cases dedicated to speech and gesture [19], speech and face recognition [11] or vision and haptics [13]. Even though the work done in tackling technical problems is of fundamental importance to the development of adaptive and multimodal interfaces, it is of a very particular nature, and not suited for a more general interface description. Also, frameworks supporting the development of interfaces for various devices exist; however,

they do not consider the specificities of multimodal interaction in its design [5][6]; or they focus only on the use of the same modality in different devices [1]; or they ignore the possibility of adapting the components properties and features in run-time placing the burden on the designer [4]. In general, they do not consider in there architectures the introduction of modalities, and how they can be explored to achieve the goals of Universal Access.

In the following, we first explain how European funded project GUIDE [7], aims to adapt interaction and UI presentation to fit each user's characteristics and level of expertise. Also, resulting from specific user trials and discussions with developers, we also show how it makes use of a User Initialization Application to know and instruct its users, and how it supports adaptation by providing developers with solutions for UI modification, and tools for helping in the development of new user-centered and accessible applications. All this attending to user needs and differences, at the same time as it takes into consideration the developer's interests.

II. CHARACTERISTICS OF GUIDE PROJECT

A. End-Users and Goals

GUIDE [7] (figure 1) aims to achieve the necessary balance between developing multimodal adaptive applications for elderly and disabled users, and preserving TV and Set-Top Box(STB) developers/manufacturers design methodologies and efforts. Consequently, there are clearly two different end-users of this project: elderly and impaired users and developers of TV based applications. Creating a bridge between these two, we have also the STB manufacturers who dictate the rules about which type and which characteristics of applications can be used on a TV based environment. Firstly, for elderly and users with impairments, GUIDE has the goal of providing new ways of interacting with a TV, by applying multimodal interaction, supporting the use of different devices as well as different combinations of input and output techniques, and adaptation to each application's UI an each user's way of interaction. In other words, elderly or impaired users who are having difficulties interacting with modern TV systems because of their complexity, will be able to interact in a more intuitive way, using alternative modalities in a single or combined fashion, while each interface characteristics will also be adapted to fit user's characteristics automatically. For all this, GUIDE has as a clear defined target environment, a STB connected to a TV in user's home (and closed) environment. Secondly because developers tend to have no concerns about accessibility when designing TV applications, GUIDE has to be capable of reducing development effort in a radical manner. For that end, GUIDE wants to create a toolbox for accessible applications and UI design, shifting the design principle from a conventional user-centered design process to a GUIDE-assisted design and development process. Through all this, GUIDE also wants to ensure that developers (and also manufacturers) can maintain the control over the modifications made on their own applications UI. Meaning, the adaptation provided by the system for adapting interfaces to user characteristics must have boundaries that cannot be crossed. And these boundaries are defined by the developers.

B. Multimodal Interfaces and Devices

Input modalities to be supported in GUIDE are based in the more natural ways of communication for humans: speech and pointing (and gestures). Complementary to these modalities, and given the TV based environment, the framework should support the usage of remote controls and other devices capable of providing haptic input or feedback. As a result, GUIDE incorporates four main types of UI components (figure 1): visual sensing and gesture interpretation; audio; remote control; haptic interfaces and a

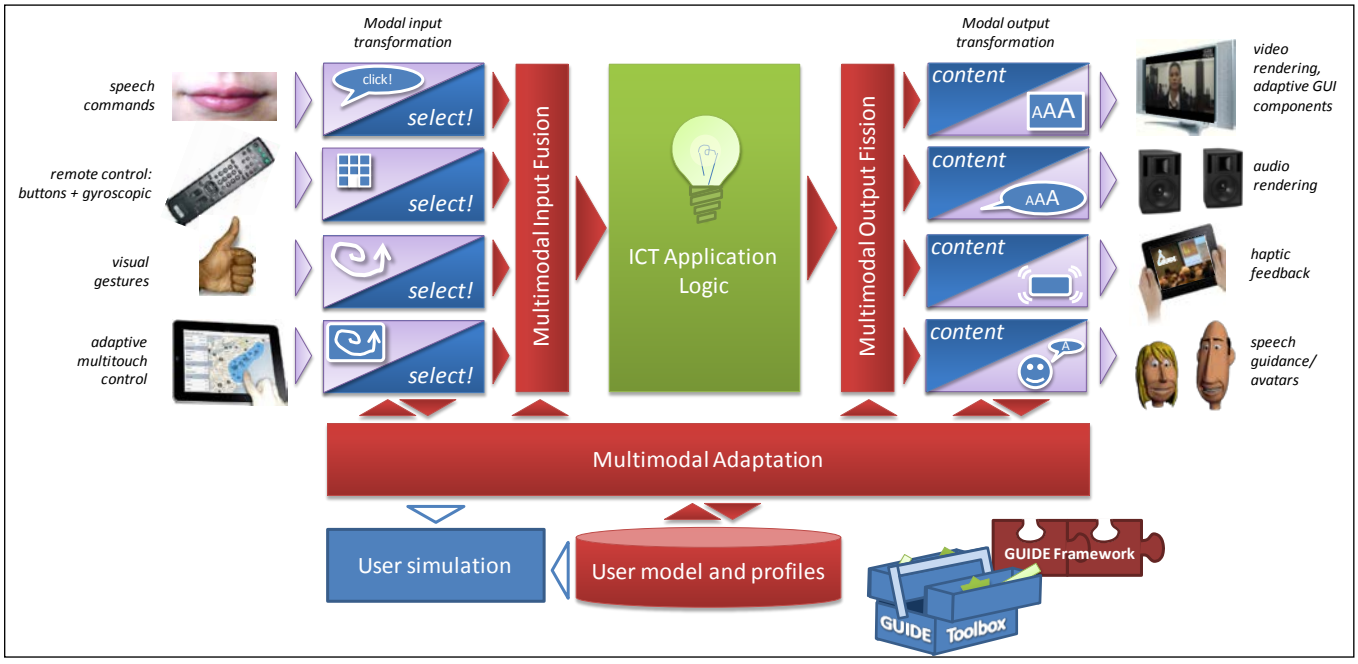


Fig 1. Features in GUIDE

multi-touch tablet. In what concerns the output modalities, the framework should consider and integrate the following output components: video rendering equipment (TV); audio rendering equipment (Speakers); tablet supporting a subset of video and audio rendering and remote control supporting a subset of audio rendering and vibration feedback. A tablet may also be used to clone the TV screen or complement information displayed on the TV screen but essentially is used as a secondary display. The main user interface should be able to generate various configurable visual elements such as text (e.g., subtitles or information data), buttons for navigation purpose, images and video (e.g., video conference or media content). Additionally also a 3D avatar is generated and expected to play a major role for elderly acceptance and adoption of the GUIDE system, being able to perform non-verbal expressions like facial expressions and gestures and giving the system a more human like communication ability.

In order for the UI to be adapted to the user's needs, these elements are necessarily highly configurable and scalable (vector-based). Size, font, location, and color are some attributes needed to maintain adaptability. These graphical elements enable the system communication with the users by illustrating, answering, suggesting, advising, helping or supporting them through their navigation. Also, both input and output modalities can be used in a combined manner to enrich interaction and reach every type of user.

C. Discussion: What GUIDE needs to know

For reaching its goals, GUIDE has to define a framework structure and collect information by asking and testing its end-users. So, the following questions have to be answered: What components the GUIDE framework has to have? What are the main preferences and typical behavior of elderly users when interacting with the system, and how to collect these preferences? How to perform automatic UI adaptation? How to help developers and manufactures in design process?

III. Learning From End Users

To get answers to the questions above, we firstly derive end user requirements from results obtained through a quantitative and qualitative analysis of data recorded in comprehensive user trials [8]. Secondly, we organized focus group sessions with developers and used an online survey as qualitative research tools in gathering additional requirements from developers and STB platform providers

A. Initial User Trials

The GUIDE project pursues a User Centered Design (UCD) process, taking into account that one of the main principles that characterize UCD is iterative design. According to this principle the system is designed, modified and repeatedly tested. This iterative cycle allows the designers to think in the product design and include the changes needed depending on the users' feedback. Following this approach, an initial study to elicit user requirements has been carried out.

1) Main Objectives

Additionally to the identification of viable usage methods (gestures, command languages) of novel traditional UI paradigms for the different impairments in the target groups via user studies in realistic user scenarios, this user trials also have the goal to generate quantitative and qualitative user data in order to establish and construct a generic user model. This user model will provide data representations for each user and will constitute the first step for adaptation in GUIDE, and will "virtualize" user impairments to try to capture the amount of knowledge needed for application design.

Fig 2. Test Application used in the technical user trials.

2) Organization and Setting

The initial user studies carried out can be divided in two different categories; one survey session and one technical trials session. While the aim of the survey was to collect qualitative information about application acceptance, user habits and modalities of interaction, the objective of the technical trials was to gather both quantitative and qualitative data and observe the interaction between the elderly and the system, performing simple tasks in the context of TV interaction. In this Test Application (figure 2), the users had the opportunity to experiment the different modalities and devices of interaction (table 1), and the tests were divided in several scripts concerning different types of interaction, or different UI elements and GUIDE aspects (table 2).

Device/Modality of Interaction	Input and Output on the User Test Application
Remote Control	Button selection (input)
Wii Remote + Wii Sensor Bar	Pointing, gesture and button selection (input)
Kinnect + Kinnect Sensor (originally a Led Camera Sensor)	Pointing, gesture and button selection (input)
Avatar Engine	Audio and visual output
Speech Synthesis	Audio output
Simulated Speech Recognition (Wizard of Oz)	Audio input
Tablet (Apple's IPad)	Touch screen input and visual and haptic output.

Table 1. User interface components used.

Type of tests	Task to perform	Devices (modalities) used
Modalities and devices experimentation	Answering to questions related with preferences of interaction, experimentation of each device and modality. Menu items selection and navigation.	Input: Remote control, Wii remote, Kinnect. Output: Visual menus and Avatar
Visual capabilities and preferences	Answering to questions related with interface visual configuration (font size and colour, background colour and button size and location tests). Menu items selection and navigation.	Input: One or more devices chosen by the user. Output: Visual menus
Audio capabilities and preferences	Answering to questions related with audio preferences (audio volume)	Input: Speech Output: Audio
Cognitive capabilities	Localization of different items on the screen (cognitive scientific tests), Measuring time of response	Input: Speech, Wii Remote, Kinnect Output: Visual menus and pictures
Motor capabilities and preferences	Performing gestures. Menu items selection and navigation. Interacting with the Tablet. Answering to questions related with motor preferences and pointing mechanisms.	Input: Wii Remote, Kinnect, Tablet Output: Visual menus
Avatar preferences	Interacting with the Avatar. Answering to questions related with Avatar preferences.	Input: One or more devices chosen by the user Output: Visual menus, Avatar
Multimodal preferences	Menu items selection and navigation. Simulation of application contexts of use. Answering to questions related with multimodal interaction and preferences.	Input: One or more devices chosen by the user. Output: Visual and Audio menus, Avatar

Table 2. Modalities, tasks and devices.

In [8] you can find a more extensive description of the tests performed.

B. Developer Focus Groups and Survey

The GUIDE system is not exclusively focused on elderly users, but also centered in developers of TV based applications and manufacturers of STBs. For this reason, major discussions regarding subjects like adaptation, elderly user's interaction, type of applications, and developers requirements for making possible the GUIDE ideas, has taken place in this evaluation, by performing both focus group with these end-users and by launching an online survey with the same user target.

1) Main Objectives

The general goal is to explore and understand the common practice among developers working on STBs. Thus the first objective is to gain data about current tools and APIs used in Set top box/connected TV platforms and to investigate how accessibility is currently perceived and applied in the industry. Secondly, exploring developer knowledge to identify which tools would developers need to efficiently integrate GUIDE-enabled accessibility features into their applications. Additionally, stimulate new ideas through discussions and to identify new relationships between objects embodying GUIDE concepts and objects embodying common practice. And finally, inform STB application development community about GUIDE.

2) Organization and setting

Developer Focus Groups: Two focus group sessions were carried out with connected TV platform providers and developers of applications and user interfaces deployed on STBs in a natural and interactive focus group setting. The sessions were conducted by two moderators (for ensuring progress and topic coverage) and each focus group session had between six and eight participants and lasted between 120 and 150 minutes. Sessions were initiated with presentations of scripts containing development and use cases that cover different aspects of the GUIDE project and its concepts. Presentations of each development case script lasted 10 minutes and were followed by 30 minutes of interactive brainstorming, and discussions.

Developer Online Survey: A questionnaire was designed to investigate how accessibility is currently perceived and applied in the industry. In addition, the survey was used as a medium to let respondents vote on the most important features of the envisaged GUIDE framework and toolbox.

Both survey and focus group were composed by the following participant types: STB test developers, STB experts in Innovative part, Flash application developers, HTML developers, middleware STB developers, architects in STB platforms, GUI developers for STB, project managers for STB projects, managers in Innovative projects for STB, product and marketing managers, research community, and standardization bodies and related organizations. In total, 81 participants from 16 countries, and 30 companies all over the world, participated.

C. Results and Conclusions

From the realization of both initial user-trials and developers focus group (and online survey), we now summarize qualitative results which will work as starting points for the next section of this paper:

1) User Survey Results

The large numbers of variables contained in the data set were submitted to a two-stage process of analysis where correlations were made and a k-mean cluster analysis [2] was performed, reducing the results to only significant data (why we used the variables listed below are described in [9]). Resulting from this, 3 user profiles capable of discriminating differences between users were created – low, medium and high. These profiles were formed by combining and grouping all modalities simultaneously such that a specific grouping may represent capability on users perceptual, cognitive and motor capability ranges. The main differences noticed were the following measures: capability to read perfectly from close and distant vision; capability of seeing at night, and color perception; capability to hear sounds of different frequencies and to distinguish conversations in a noisy background; cognitive impairments; and mobility diagnosis like muscular weakness and tremors. Table 3 shows all the identified variables in the three profiles.

Vision	LOW	MED	HI
Close vision: level able to read perfectly	20/20	20/60	20/80
Distant Vision: level able to read perfectly (metres)	5	5	20
general eyesight	good	excellent	normal
seeing at distance	good	poor	poor
seeing at night	normal	poor	poor
colour perception	good	bad	bad
Hearing	LOW	MED	HI
Able to hear a sound of 500Hz?	Yes	Yes	No
Able to hear a sound of 2Khz?	Yes	Yes	Yes
conversation from a noisy background	excellent	normal	normal
Cognition	LOW	MED	HI
TMT (seconds) Cognitive executive function	30 (no impairment)	49 (low impairment)	136 (high impairment)
Motor	LOW	MED	HI
mobility diagnosis	none	hernia / slipped disc	none
muscular weakness	never	A few occasions	Frequently
write	No difficulty	No difficulty	Mild difficulty
Tingling of limb difficulty	No	Mild	Mild
Rigidity difficulty	No	Mild	Moderate

Table 3. Cluster centers

2) Technical User Trials Results

Big, centered and well-spaced buttons were preferred by users because they are easier to see and select (and elderly users typically have some kind of visual and motor impairments). Additionally, users prefer medium sized fonts and medium volumes for audio, but users with impairments tend to prefer bigger fonts and higher volumes. However, more than based on user abilities or preferences, both visual and audio elements configuration, depends on the interaction context and must be at all times modifiable and repeatable by the user. All the preferences described regarding visual components, reflect the low efficiency (lot of time needed for each selection) and accuracy (wrong target when selecting) registered when interacting with any type of pointing in these tests.

Users clearly preferred gestures easier to make (swipe and pinch), and have no problem whatsoever interacting by gestures. It was also evident that alternative ways of interacting with the TV (speech and finger pointing) are preferred to the traditional way. Also, training makes any type of modality more efficient as the user learns what is required to perform each interaction. However, any type of interaction should not be imposed on the users, but be available as an intuitive option for interacting with the TV. Additionally, when not used by the user intuition, modalities of interaction should be explained to the user before he or she starts using the system.

Every user is able to interact multimodally with the system and combine speech and pointing, even when they prefer only one modality. Users exhibited different multimodal interaction patterns during the trials and there is no specific interaction pattern for each user (a user can speak first and point afterwards, and in the next interaction do the opposite). Users can also change the way they interact depending on the type of feedback given while interacting. Regarding user preferences in input and output modalities, there are clear differences between what users say they prefer, and what users really ask for when interacting. In fact, 100% of the users want multimodal output every time information is presented to them, because every user who said to prefer only one type of feedback admitted differently when in specific interaction contexts. The same happened concerning input modalities, with almost half of the users admitting, when confronted with practical tasks, that they were wrong when they said to prefer only one modality.

The results obtained in these trials enforce the need of a multimodal system and also the need for adaptation, as we can see in a more detailed fashion in [8].

3) Developer Focus Groups and Survey Results

Developers agree that if users are involved in every development phase of the applications (or in the maximum phases possible), the resulting UI will be more usable. It was concluded that for elderly people UIs should be maintained clear and simple, however without giving the impression that it has been designed for someone with impairments (not leaving the feeling of a “system for seniors”). Additionally, costs are the current major reason for reduced application of user-centered design in the industry (followed by time and

lack of awareness). As the current most important device on interaction with STBs, the remote control must continue to have a central role in the interaction, and should only be relegated to a secondary role if that is a result of each user interaction preferences. Gesture control and speech input are recognized as secondary technologies. In general, participants agree that automatic adaptation of user interfaces can help elderly users to access ICT services. However GUIDE adaptation mechanism should never change interface aspects unless it is mandatory for specific user interaction. Also, radical changes in the UI must be avoid so that the user feels he/she is in control and not get lost in the interface. If a radical change is indispensable the UI must inform the user of the proposed changes. Identified as the main obstacle to UI adaptation is the fact that elderly users present too many differences between each other. Therefore, for adaptation to fit each user, GUIDE has to first find a way to know his or her impairments, preferences or characteristics. This "discovery" will have to occur the first time the user interacts with the system, and will have to be short, not too much intrusive and entertaining to the user. The most important conclusion debated in this subject is the one saying GUIDE should support UI mark-up as interface between application and GUIDE adaptation. This way, developers will be allowed to keep tools and development environments and without too much additional effort, take a first step to accessible design. Web developers mostly use HTML editors as the most important tools in Web & TV development. However, having to learn new development processes will drive developers away from the GUIDE framework. So, developers should not be required to develop taken into consideration specificities of the multimodal operations but have a clear specification of how such devices interact with the framework. As it was already described in UI adaptation results, identification of UI components should be made using only mark-up language, however applications coded using dynamic HTML (through JavaScript) must continue to be able to change, remove or insert elements in the currently rendering page. Meaning, all changes in application presentation will need to be identified at run-time. For most participants connected TV platforms and STBs will be most relevant platforms in the future. Also, Web-based application environments will become more important for Web & TV. Manufactures stated increasing STBs capabilities cannot raise its price to much, or development will be more difficult and costly. Developers also pointed out GUIDE system must consider situations where multiple users are using the TV and services.

IV. MULTIMODAL APPLICATION DEVELOPMENT

From the results and implications reported in the previous section of this paper, we now derive GUIDE project solutions for giving answers to the same questions raised in the beginning of this paper.

A. *Multimodal and Adaptive Framework*

We now give an overview of the GUIDE framework [8] (figure 3) following an interaction cycle, starting from the

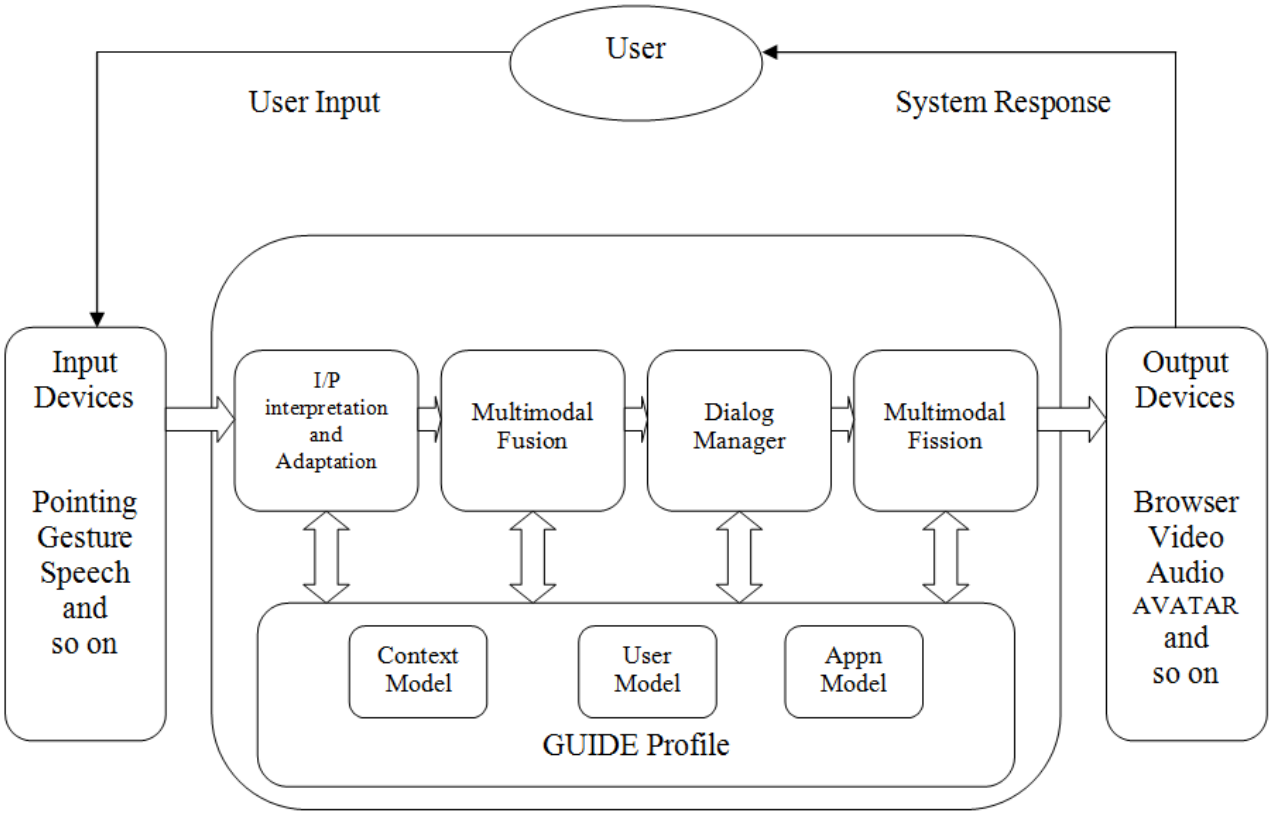


Fig 3. Multimodal and Adaptive Framework architecture in GUIDE

user input and going through the construction of the system's output to be presented to the user.

A user provides input through multiple devices and modalities which can be used simultaneously. The signals from recognition based modalities are processed by interpreter modules (e.g., a series of points from the motion sensor go through a gesture recognition engine in order to detect gestures). The signals from pointing modalities go through input adaptation modules (e.g., in order to smooth tremors from the user's hand). Both interpreter and adaptation modules base their decisions on knowledge stored in the user profile, thus improving the efficiency of noise reduction in the input signals. Then, the multimodal fusion module receives, analyses and combines these multiple streams (outputs of input interpreters and input adaptation modules, or raw data that did not go through any of these) into a single interpretation of the user command based on the user, context and application models (abstract representation of the application). This interpretation is sent to the dialogue manager who decides which will be the application's response, basing its decision on knowledge about the current application state and the possible actions that can be performed on the application in that state. The dialogue manager decision is fed to the multimodal fission module, which is responsible for rendering a presentation in accordance to which output to present (derived from the application itself and the application model), the user abilities (accessed through the user model) and the interaction context (made available through the context model). The fission module takes all this information and prepares the content to render, selects the appropriate output channels and handles the synchronization, both in time and space, between channels when rendering. This rendering is then perceived by the user, which reacts to it, and starts a new cycle by providing some new input.

B. *User Initialization Application*

In both technical user-trials and focus groups, it is the necessity of knowing every user characteristics, preferences and impairments from the first time he or she interacts with the system. This is mandatory because of the user's differences and the necessity of adapting both UI components and interaction to fit each user, as well as the necessity of instructing the user about every possibility of interaction in order to reach the maximum efficiency when using the system. GUIDE adaptation begins through a User Initialization Application (UIA) (figure 4) that allows for the acquisition of primary assumptions about the user. So, knowing that each user model contain assumptions about interesting characteristics of user subgroups, after "going trough" the UIA, a user is assigned to a user model as certain preconditions are met. From that moment on, and for any GUIDE application the user interacts with, the system is "initially" adapted to him/her. It's relevant to say that the UIA is presented to the user as a simple step-by-step configuration of a "general" interface. In each step, different types of contents and different contexts of interaction are presented, so the user can test different components and parameters, and the system learns the user characteristics, from his impairments to his preferences. Addressing the results from the developer focus groups, every UIA run as to be short in time, intuitive and transparent to the user and also

Fig 4. Screenshots of the first version of the User Initialization Application

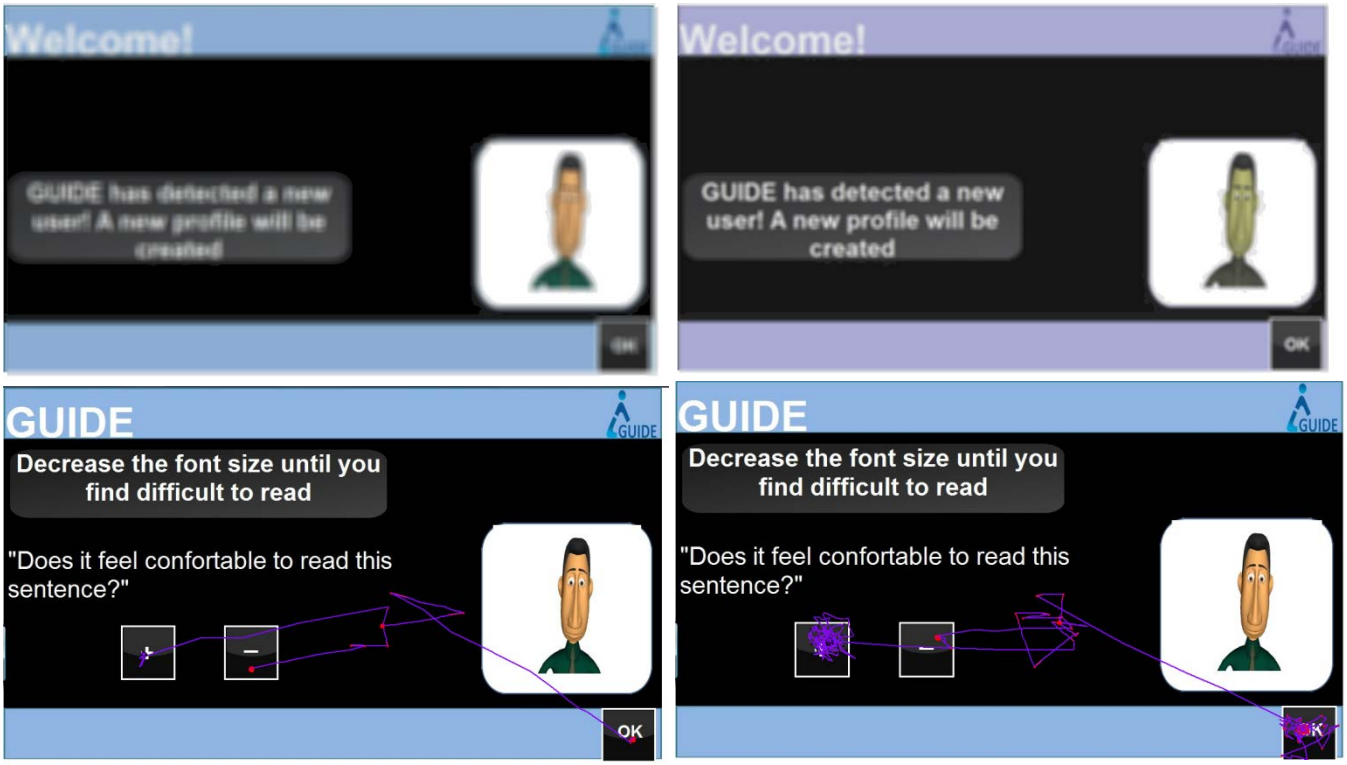


Fig 5. Applying simulation of User Initialization Application (top images show visual impairments - Macular Degeneration and Visual Acuity Loss -, bottom images show motor impairments - moderate motor impairment and Parkinson's Disease)

serve as a "tutorial" for learning every modality of interaction available in the system. Additionally, the user must be recognized (facial or voice patterns) by the system so that the information provided can be stored in a profile and loaded every time the user interacts with the system.

C. *Simulation of User Impairments*

As developers need tools for helping saving time and cost in the development of inclusive TV base applications, GUIDE offers a simulator [2] which will allow the developer to perform accessibility tests based on virtual users, saving much time in comparison to tests with real users. So, evaluation as a typically expensive step in user centered design is supported in GUIDE by a simulation functionality allowing to illustrate to developers how users with typical impairment profiles will perceive or may interact with an application. The simulator can show how certain visual and strength impairments influence the way a user perceives and visualizes a certain UI (e.g., how an elderly color blind user sees a specific UI), and also what are the effects of those impairments in the user interaction (e.g., predicting cursor paths on the screen or task completion times). This simulator can be characterized as a tool for helping developers to take adaptation into consideration at design time. Figure 5, shows the simulation results on top of the User Initialization Application for both visual and motor impairments.

D. *Filtering*

As verified by the inefficient and erroneous use of pointing interaction when performing selections in the user trials, elderly users potentially have a wide range of impairments that hinder their ability to communicate their intentions to an application. In some cases these impairments can be severe, and significantly affect the speed and accuracy. This leads to an inefficient or even undesirable interaction with an application. The use of cursor smoothing techniques in GUIDE consists in processing the raw user input to obtain a filtered input (Input Adaptation Module described in the framework). This requires the usage of efficient statistical signal processing schemes to estimate the user's intended operations in real time. Basically it consists in the application of corrective forces and forcing relatively smooth paths in a cursor interaction as well as assigning attraction fields to UI elements. Therefore, the following graphical UI filters can help improving pointing interaction within the GUIDE project:

- Exponential averaging: this modification calculates the cursor position pi as $pi = \alpha xi + (1-\alpha)pi-1$, where xi is the user input, pi-1 is the previous cursor position and $\alpha \in [0,1]$ is a parameter determining how strong the user input influences the cursor position. This method produces smooth cursor traces but has the drawback that it can produce a delay between user's intended position and the actual position;

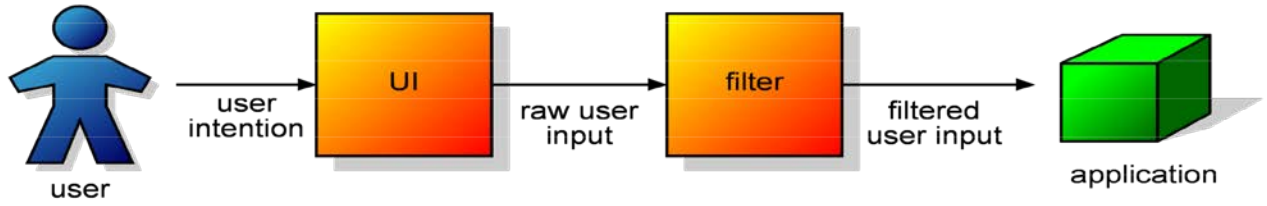


Fig 6. Process of detection of user intentionality by applying filtering

Fig 7. An example of missed clicking (left) and clicking with the gravity well filter (right)

- Damping: This method introduces a quadratic force that opposes the velocity of the cursor preventing sudden changes in directory or speed when interacting;
- Gravity well: This method warps the cursor space, generating attractive basins to ease the selection of visual targets. This simplifies pointing interaction selection forcing the selection of buttons or UI elements that are more close to the location where the user is pointing (figure 7).

Considering the different user characteristics and impairments, and the different UI element configuration, the existence of these filters make possible that motor impaired users can more easily interact with pointing and also makes possible the use of small and less spaced buttons in applications UIs avoiding errors in selection caused by the proximity of the buttons. All because pointing interaction accuracy is raised, and user intentionality when pointing is taken into consideration (figure 6).

E. Semantic Programming and Run-Time Adaptation:

The specification of TV based applications in GUIDE will be based on Web-based languages like HTML, CSS and JavaScript because of their wide acceptance among developers and compliance with STB specifications. However, in GUIDE exists the additional side-condition of specifying multimodal applications that needs to be merged with these web-based specification languages. This is made by specifying additional information about how an application is supposed to adapt in different modalities. For this semantic annotations are added to the HTML code, based on the WAI-ARIA draft specification of the W3C.Only by providing this type of supplementary information it is possible for the system to create an abstract representation of the application. Then, using an automatic application transformation module the system converts the annotated application description into a modality-independent application representation, the Application Model described in the framework. Subsequently, and depending on the user interacting and on the level of control defined by the application developer, adaptation of UI components is performed.

Developers can create their applications and UIs in an established manner, and GUIDE automatically adapts the UI to the user. This avoids having to design many user interface templates for various heterogeneous user groups. Therefore, GUIDE provides the application developers with two possible levels of adaptive control:

Augmentation: presentation and interaction options taken by the developer are not subject of change. Instead, if the user model suggests that the presentation is insufficient for the user abilities, the presentation is augmented in different modalities (for example supplementing a visual interface with sound feedback). The multimodal fission mechanism renders the application output directly, augmenting or not the rendered presentation depending on the user model. Figure 8 shows an augmentation example;

Adjustment: application rendering is adjusted to the abilities of the user (for example adjusting components of a visual interface to fit user characteristics, like raising font size or button size). The rendering changes can be achieved

through CSS manipulation. Adjustment can be combined with augmentation. Figure 9 shows an adaptation example.

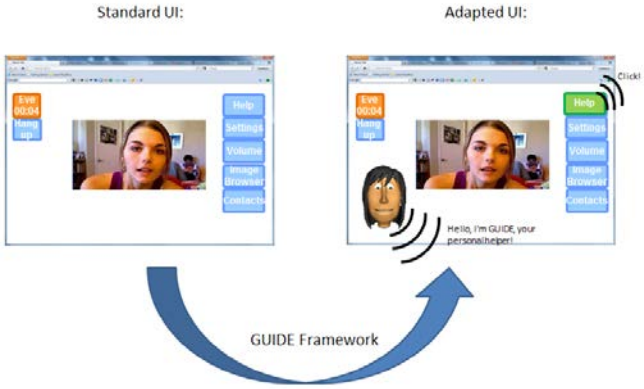


Fig 8. Example of Augmented adaptation

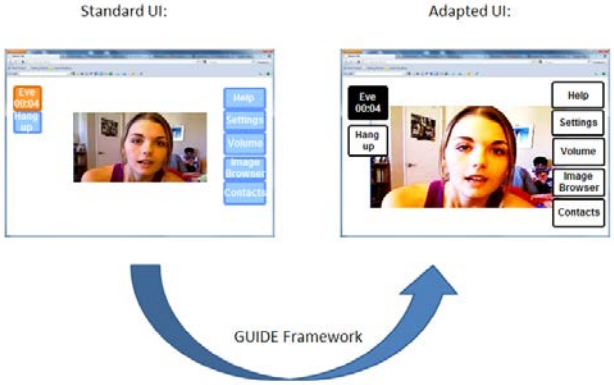


Fig 9. Example of Adjusted adaptation

V. DEVELOPING AND EVALUATING THE USER INITIALIZATION APPLICATION

A. Development

1) Selection of tasks and metrics

The tasks and metrics chosen for the UIA are the ones for which the resulting data is the most capable to assign the more appropriate profile to the user profile. They were selected from an analysis of the extensive survey data, taking into account the feasibility of gathering the data. For those instances where it was not feasible to gather the data in a living room environment, alternative sources were selected and combined to estimate the required data. A description of these variables is listed below: **Color Blindness**: Plates 16 and 17 of Ishihara Test [6] as it may classify among Protanopia, Deuteranopia and any other type of color blindness; **Dexterity**: We estimated Grip Strength and Active Range of Motion of wrist from age, sex and height of users following earlier Ergonomics research [2]; **Tremor**: We conducted earlier a test involving a Tablet device in horizontal position, and estimated tremor from the average number of times users need to touch the screen to select small buttons. Details of the study can be found in a separate paper [4]. Additionally, other tasks were chosen with the purpose of allowing users to personalize the system, while being a hands-on tutorial regarding new modality interaction and feedback configuration. The most relevant ones are the following: **Modality Introduction**: Self-explanatory videos of how to interact with each modality, followed by "do-it-yourself" tasks; **Button and Menu Configuration**: Button size, and font and background color configuration; **Cursor Configuration**: Cursor size, shape and color configuration; **Audio Perception**: Hearing capabilities and preferences.

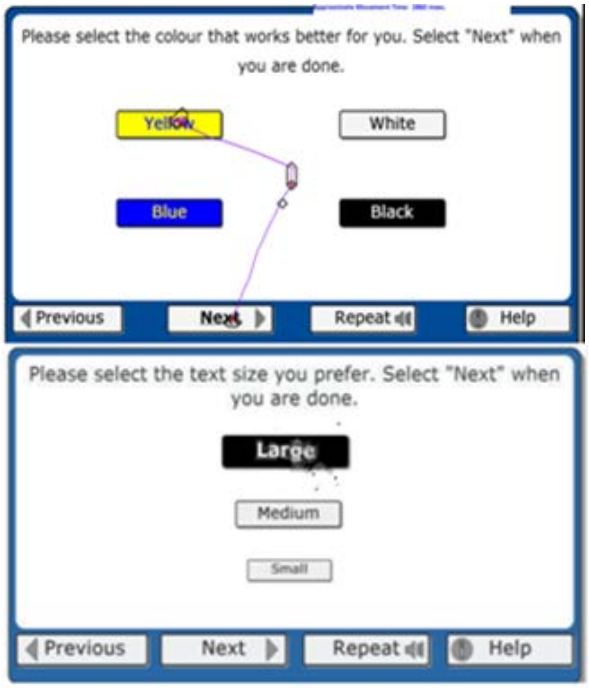


Fig 10. Applying Simulation on the UIA

2) Accessible Interface

The UIA has a simple user interface, with a different screen for every task and metric identified above. Few buttons are presented per screen (preventing user confusion). Every screen preserves the same navigation model - an area with "next", "previous" and "repeat" buttons, and another visually distinct area for presenting information and requests. For every metric to be measured, tests are presented as simple questions about preferences. Also, for every modality available in the system, a video introducing its use is presented, followed by the possibility for the user to try it out. A virtual character (Figure 11. first screen) accompanies the user through this process, offering explanations and assisting the user in the personalization. As the user goes through each task and preference setting, the UIA adapts itself to the preferences already manifested. For example, if user manifests preference for big, blue buttons with yellow text, all buttons will be presented with those settings from that moment onwards. It is worth pointing out that the results of our previous study are reflected in the UIA's design: high contrast colors, big, centered and well-spaced buttons, etc. Also, the GUIDE simulator was also used in the design of the UIA, to ensure that users with visual and motion impairments could use it with high efficiency (figure 10).

B. Evaluation

1) Study description

With the goal of evaluating the efficiency and acceptance of the User Initialization Application by elderly a study was conducted. First we want to measure the efficacy of this application in discovering the relevant characteristics of users and assigning user profiles; and secondly, we want to evaluate how understandable the UIA is in terms of its goals and the instructions it provides; and finally, how easy it is for elderly to interact with this application, or if they would do it if it was part of their daily lives.

2) Participants (Pre-Survey)

We recruited 40 elderly people (24 female and 16 male) with different age-related disabilities. Users were recruited in two countries, with 21 participants (14 female and 7 male) being recruited in Spain and 19 participants (10 female and 9 male) in the UK. The average age was 70.9 years old and the different user profiles were assigned to the participants in the following manner: 14 users with profile A, 22 users with profile B, and 4 users with profile C. All users participated voluntarily and all activities involved in this study were safeguarded from the ethical point of view.

3) Apparatus

The study was conducted in two locations (Spain and UK). Efforts were directed to create similar environment and technical conditions in both labs. Trials were conducted by usability experts. Users were given freedom to interact (the trial conductor would only intervene when really needed, or user asked for help). In what concerns the technical setup and specification, different modalities of interaction were configured: pointing resorted to the use of a Microsoft Kinect; for speech recognition we used the Loquendo SR engine; a simplified remote control, with less buttons than traditional ones and capable of controlling pointer coordinates using a gyroscopic sensor was made available; an iPad was used for tablet interaction; and a full 1080p HDMI TV with integrated speakers and a 32'' screen was used for visual and audio output. User interactions and answers were video recorded.

4) Design and Analysis

We used a within-subjects design where all users ran the UIA. Qualitative analysis was retrieved from pre, intermediate and post-questionnaires. Quantitative data was retrieved from the UIA (user profile and interface preferences). Herein, we discarded quantitative measures like trial errors and time as the trials followed a semi-supervised methodology: the participants were motivated to perform the tasks on their own but they were free to ask questions when they felt lost. For binomial measures, Mcnemar's test was performed, and Cohen's Kappa was used to assess the inter-reliability of the profile ratings.

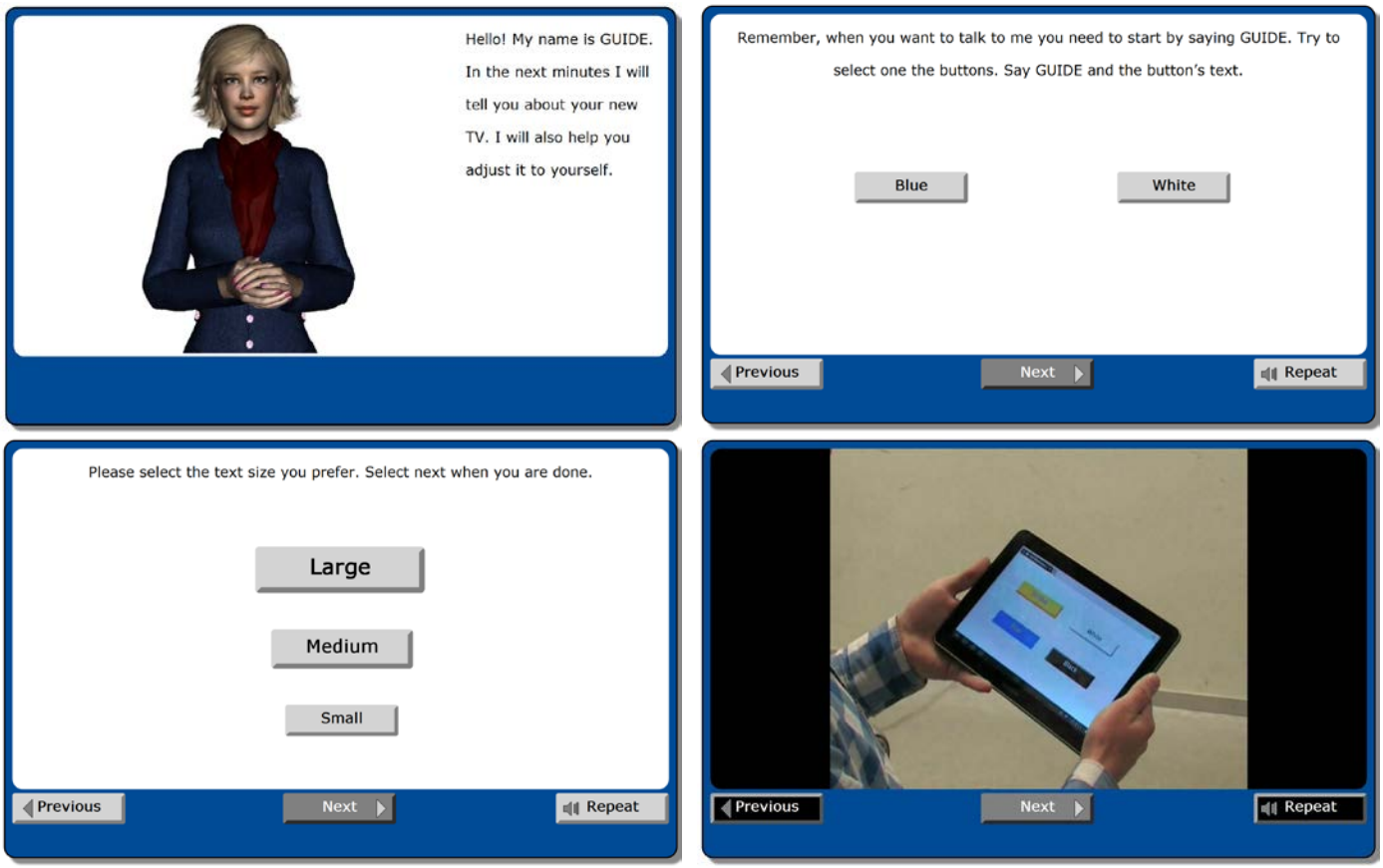


Fig 11. Implemented Version of the User Initialization Application

C. Results

1) Discovering Elderly Profiles with UIA

Our take for adaptation relies on a User Model fed by the UIA. All participants in our study performed both the pre-survey and the UIA. Twenty-nine out of forty profile assessments were performed similarly by the two methods (74%). The interrater reliability between the profiles assigned with the pre-survey and the UIA was found to be Kappa = 0.58 ($p < .0.001$), revealing a moderate agreement [21]. It is relevant to notice that the UIA enables the user to input preference values, something that goes beyond ability profile. This is likely to explain part of the mismatch (e.g., a user with no visual impairments is likely to prefer a higher contrast button when he is confronted with such an hypothesis). Another source of uncertainty may be the under-statements by part of the users in the pre-survey. Indeed, in a questionnaire it is likely that part of the users fail to acknowledge some limitations while they clearly state them when confronted with an interface with options to surpass it. A deeper understanding of the mismatches that are not created by these observed flaws can only be retrieved in a more extensive evaluation by analyzing how both methodologies enable the users to improve performance.

2) UIA evaluation by the Elderly

Users are not used to use something like UIA, so it is important to assess how the users see this component and if they are willing to use such a thing to improve their performance.

Question about the UIA	Median	IQR
Have you understood why we do the UIA? [1 - Yes ; 2 - No]	1	0
If you have had the system at home, would you go through it or skip it? [1 - Would do it ; 2 - Would skip it]	1	1
Do you think the UIA is too long? [1- Yes;2 - Neutral;3 - No]	3	0
Were the instructions easy enough to understand? [1 - Yes; 2 - No]	1	0
Did you notice any changes in the application while you were using it? [1 - Yes ; 2 - No]	1	1

Table 4: Subjective ratings to the UIA

The participants took between 12 and 37 minutes to complete the UIA (M=22.8, SD=5.9). Once again, although they were discouraged to engage in long dialogues the participants were free to express their opinions and doubts during the UIA which increased the time to finalize the process. The UIA classified 16 people as profile A, 20 as profile B, and 4 as profile C. Table 4 presents the subjective ratings given by all the participants to the questions posed. Regarding the understanding of the purpose of the UIA (Question 1), 9 out of 40 (22%) did not understand the purpose of the UIA. This indicates that such a process should be better motivated or else it will be likely ignored by the users. In line with this, 11 out of 40 (28%) stated they would skip the process if they had the system at home (Q2). Five participants stated to find the process too long while four other were neutral about it (Q3) All the remaining thought it was neither too long nor tiring. Most users (35) thought the UIA was easy to follow and understand (Q4). Regarding the adaptations felt during the UIA (Q5), 26 participants stated to have noticed them. This is easily explained as 16 participants were classified as profile A which means they had little or no adaptations done during the UIA. In sum, the users seem positive towards the UIA (Table 2) although it is clear that it should be well motivated and accompanied.

D. Discussion

Upon analyzing the UIA process and its impact on adaptation along with the usage of the GUIDE system and its underlying concepts, we answer our research topics as follows:

1) Deriving a suitable user adaptation profile through the UIA.

The UIA aims at creating a user profile by performing a simple set of questions and interactive tests. Results showed that the UIA is able to match profiles obtained with an extensive survey in 74% of the cases. Further, the UIA showed to be more realistic than its paper-based counterpart as data is likely to be more accurate when the users are faced with their limitations rather than just being questioned about them. Moreover, the UIA gives space for preference and subjectiveness. In sum, we consider that adapted TV applications based on simple initialization profiling are feasible and likely to improve over traditional methodologies.

2) Acceptance of the UIA.

The UIA took over 12 minutes, averaging around 23 minutes. This amount of time can be discouraging for an elderly user if the benefits are not clear. Taking in consideration that it is supposed to be ran only once, the participants showed to be very positive about it. This is supported by the almost general understanding of the purpose of the UIA: they understood the benefits of such an application and perceived the adaptations during the process. Most participants (35) considered the application easy to follow which indicates that although the concepts underlying the creation of the user model are complex, the interface to generate it is not.

VI. CONCLUSIONS

New interaction paradigms, supported by new modalities and applications, are transforming a classical appliance that is the TV. If not handled properly, this transformation can increase the access barriers to TV content for elderly users.

In this paper, we assessed several of the proposals that the GUIDE project puts forward in order to increase the accessibility of TV applications. GUIDE aims to provide application developers with a multimodal adaptive framework and a set of functionalities that will increase their products' accessibility, without demanding major changes in their development process. The assessment was based on a user trial, with 40 participants from two different countries.

The results obtained in this technical user-trials about the existence of disparity between what modalities users say they

need, and what modalities they ask for when using the system, favors multimodality almost every time. This only helps to prove that the use of several input and output modalities is indispensable in the development of multimodal TV based applications for all. Also indispensable, are the components identified in the GUIDE framework, and the combined use of semantic programming and run-time adaptation mechanisms to fit UI components to each user characteristics. Additionally, the use of a simulator of user impairments can help developers understand at design-time how certain UI templates and components are perceived by different users with different impairments, preventing user exclusion and making accessible applications easier to design.

Being an adaptive framework, it relies on knowledge and information about users. Essential for both providing and collecting knowledge, is the UIA, a process that streamlines user profile identification, based on short number of tasks and questions. We present an assessment of the efficacy of this process, concluding that it is possible to reliably identify user profiles, while also recognizing ways in which to further improve the process. From the user's point of view, the process motivation was understood, and it was considered easy enough, although also here we were able to find ways to improve it.

These results show a positive acceptance of the GUIDE concepts and their expected impact in the quality of life of its users, validate the approach followed so far and pave the road for the project's future developments, which will be verified in a longitudinal trial for better assessing the effects of adaptation and multimodality.

A. *Future work.*

Regarding the use of modalities, speech interaction was singled out as the most attractive modality. In this study, a Wizard-of-Oz approach was used to replace the speech recognition engine, and as we question ourselves on how that might have contributed to the results, a follow-up study, where a real speech recognition engine is used, is necessary. It seems safe to say that speech plays an important role in promoting the adoption of these systems, and efforts to ensure its adequate operation are justified by the satisfaction it provides users with. Tablets, although not fully integrated with the system in this study, collected a positive response from participants, with 92% of them considering interacting with a TV using the Tablet. This tendency is also to be confirmed in the future with a study where users may be asked to execute tv-related tasks on a tablet. Finally, regarding the clustering process, by increasing the number of users available it will be possible to update the profiling process, resulting in a more accurate representation of the users' characteristics and a more precise identification of the relative importance of each variable.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement nº 248893.

REFERENCES

[1] Coelho, J., Duarte, C., Feiteira, P., Costa, D. and Costa, D.. Building Bridges Between Elderly and TV Application Developers. In ACHI 2012

[2] Balme, L., Demeure, A., Barralon, and N., Coutaz, J. & Calvary, G.. CAMELEON-RT: A Software Architecture Reference Model for Distributed, Migratable, and Plastic User Interfaces. In EUSAI'2004

[3] Biswas, P., Robinson, P., and Langdon, P.: Designing inclusive interfaces through user modelling and simulation. International Journal of Human Computer Interaction.

[4] Blechschmitt,E., and Strodecke,C.: An architecture to provide adaptive, synchronized and multimodal human computer interaction. In MULTIMEDIA '02, NY, USA, pp. 287-290.

[5] Bouchet, J., and Nigay, L.: ICARE: a component-based approach for the design and development of multimodal interfaces. In CHI '04, NY, USA, pp. 1325-1328.

[6] Calvary, G., Coutaz, J., and Thevenin, D.: A unifying reference framework for the development of plastic user interfaces. In EHCI '01, London. UK, pp. 173—192.

[7] Calvary, G., Coutaz, J., Thevenin, D., Limbourg, Q., Souchon, N.,Bouillon, L., Florins, M., and Vanderdonckt, J. (2002). Plasticity of user interfaces: A revised reference framework. In TAMODIA '02, Bucharest, pp. 127-134.

[8] Coelho, J., and Duarte, C.: The Contribution of Multimodal Adaptation Techniques to the GUIDE Interface. HCII2011, Orlando, Florida, USA, pp. 337-346.

[9] Coelho, J., and Duarte, C., Biswas, P., Langdon, P. Developing Accessible TV Applications, Proceedings of ASSETS 2011, pp. 131-138.

[10] Dragicevic, P., and Fekete, J.D.: The input configurator toolkit: towards high input adaptability in interactive applications. In AVI '04, ACM Press, NY, USA, pp. 244-247.

[11] Elting, C., Rapp, S., Mohler, G., and Strube, M.: Architecture and implementation of multimodal plug and play. In ICMI '03, ACM Press, NY, USA, pp. 93-100.

[12] Garg, A., Pavlovi´ c, V., and Rehg, J. (2003). Boosted learning in dynamic bayesian networks for multimodal speaker detection. Proceedings of IEEE 91, pp. 1355–1369.

[13] Gotz, D., and Mayer-Patel, K.: A general framework for multidimensional adaptation. In MULTIMEDIA'04, NY, USA, pp. 612-619.

[14] Harders, M., and Szekely, G. (2003). Enhancing human-computer interaction in medical segmentation. Proceedings of IEEE, 91, pp. 1430–1442.

[15] Martin, J.C., Julia, L., and Cheyer, A.:. A theoretical framework for multimodal user studies. In CMC98, Tilbur, Netherlands, pp. 104-110.

[16] Oakley, I., Brewster, S. A., and Gray, P. D.: Solving multi-target haptic problems in menu interaction. CHI'01, Seattle, USA, pp. 357-358.

[17] Oviatt S. L. Multimodal interactive maps: Designing for human performance. Human-Computer Interaction, 1997, pp. 93-129

[18] Oviatt, S. L., DeAngeli, A., and Kuhn, K. Integration and synchronization of input modes during multimodal human-computer interaction. CHI '97, New York, USA, pp. 415-422.

[19] Oviatt, S.L.: Mutual Disambiguation of Recognition Errors in a Multimodal Architecture. CHI'99, Pittsburgh, USA, pp. 576-583.

[20] Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, ICai, G., Brewer, I., Maceachren, A.M., and Sengupta, K. (2003). Speech-gesture driven multimodal interfaces for crisis management. Proceedings of IEEE, 91, pp. 1327-1354

[21] Vitense, H. S., Jacko, J. A., and Emery, V. K. Multimodal feedback: An assessment of performance and mental workload. Ergonomics 46, 2003, pp. 66-87.

The impact of workload on energy efficiency of virtualized systems

Jukka Kommeri
Helsinki Institute of Physics, Technology program, CERN, CH-1211 Geneva 23, Switzerland
jukka.kommeri@cern.ch

Tapio Niemi
Helsinki Institute of Physics, Technology program, CERN, CH-1211Geneva 23, Switzerland
tapio.niemi@cern.ch

Olli Helin
Helsinki Institute of Physics, Technology program, CERN, CH-1211 Geneva 23, Switzerland
olli.helin@cern.ch

***Abstract*—Virtualization, i.e., running several virtual computers on the same physical hardware, is an essential technology in data centers. Since demand for cloud computing services is constantly growing, an increasing number of data centers are focusing on improving their energy efficiency. This has made energy efficiency of virtualization technologies an important research domain. So far, maximizing performance of virtualization technologies has received a lot of attention in cloud computing industry and several academic studies on performance optimization can be found, too. However, these studies usually focus on improving energy efficiency by applying server consolidation methods. In this paper we focus on energy efficiency of virtualization technologies, i.e., how a virtual service can be made more energy efficient. Our aim is to reduce energy consumption without decreasing the quality of service. We have studied this by performing a large set of measurements with different system settings. We used both synthetic benchmarks and real applications. We found out that energy efficiency depends on 1) the workload of the virtual servers, and 2) the number of virtual servers on the physical server. We noticed that it is more energy efficient to maximize workload of virtual servers and to minimize their number. Additionally, we observed that properly configured idle virtual servers hardly increase energy consumption. Thus, our conclusion is that it is better to load virtual servers heavily or let them run idle.**

Keywords-virtualization; energy-efficiency; server consolidation; xen; kvm; invenio; cmssw

I. Introduction

The work presented in this paper is based on our earlier work, that was published in ENERGY 2012 conference [1], and partially also on work, that was published in ICGREEN 2012 conference [2]. The current paper contains enhanced background discussion and more detailed analysis on results and also presents some new results such as latency measurements.

Web based applications have gained popularity and an increasing number of these applications are hosted by cloud computing in large data centers containing thousands of virtualized servers [3], [4]. Traditionally, a server has been purchased to host only one service (e.g., a web server, a DNS server). This is not very efficient, since according to many studies the average utilization rate of a physical server hosting a web site is around 15% of maximum but depends a lot on the service and it can be even as low as 5% [5], [6].

This level of utilization is very low compared to any field in industry. A common explanation for the low utilization is that data centers are build to manage peak loads. However, this is not a new data center specific issue, since high peak loads are common in many other fields. Even with this low level of utilization the servers are usually operational and consuming around 60% of their peak power [7]. Low utilization level is inefficient through the increased impact on infrastructure, maintenance and hardware costs. For example, low utilization reduces the efficiency of power supplies [8] causing over 10% losses in power distribution. Thus, servers should run in near full power when they do value adding work, because then they operate most efficiently considering consumed energy per executed task [5].

Scientific computing clusters at CERN have traditionally allocated resources for one analysis job such that the job gets one computing core and 2 GB of memory. As the number of cores in the CPU and the number of CPUs in the server increase, more jobs can be processed in parallel by the server. Modern servers for scientific computing can have 16 cores per CPU, two to four CPUs and hundreds of gigabytes of memory. Combining this with the need for different analysis environments, computing resources should be divided into smaller logical units.

Server consolidation by using virtualization technologies is a solution for increasing utilization, since it allows one to combine several services into one physical server. In this way, these technologies make it possible to take better advantage of hardware resources. Virtualization makes it possible to create logical containers, virtual machines, that contain a complete operating system with a user specific analysis environment. This virtual machine can be modified to meet users resource requirements and moved between physical machines to improve the total energy efficiency of a larger server cluster or computing center [9].

In this study, we focus on energy efficiency of different virtualization technologies. Our aim is to help the system administrator to decide how services should be consolidated to minimize energy consumption without violating the quality of service agreements. Virtualization in the context of this

paper refers to system virtualization where several operating system instances are run on single physical hardware, as depicted in Figure 1.

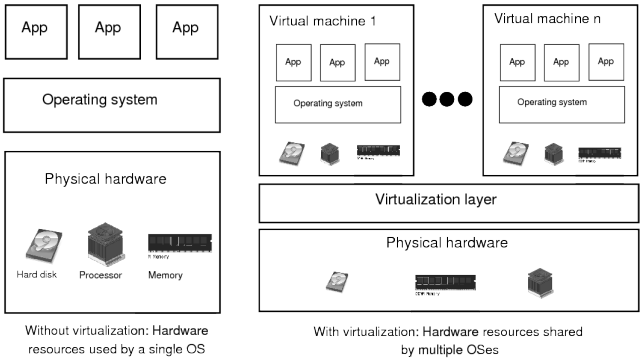


Figure 1. A non-virtualized and a virtualized system

We studied energy consumption of virtualized servers with two open source virtualization solutions; KVM and Xen. They were tested both under heavy load and being idle. Several synthetic tests were used to measure the overhead of virtualization on different server components. Also two realistic test applications were used: 1) the Invenio catalog program's database service, and 2) CMSSW, the CMS Software Framework, a physics analysis software for the data generated by the Compact Muon Solenoid, CMS, experiment at CERN. The results were compared with the results of the same tests run directly on hardware without any virtualization layer. We also studied how overhead of virtualization develops by sharing resources of physical machines equally among different number of virtual machines and running the same test set on each virtual machine set.

The paper has been organised as follows. After introduction, we review the related work in Section II. Our test environment and tests are explained in Section III and their results given in Section IV. Finally, conclusions are given in Section V.

II. Related work

Virtualization and its performance is a well-studied area. Previous studies mainly focus on performance, isolation, and scheduling. Even though energy efficiency is one of the main reasons for server consolidation and virtualization, it has not received much attention. Instead, many of the existing studies evaluate overhead differences between different virtualization solutions and how virtual resource could be provisioned between physical servers in an energy-efficient way.

Virtualization technologies are a key component of cloud computing [10]. Large data centers host cloud applications on thousands of servers [3], [4]. In such environments, the benefits of virtualization are obvious. Xu et al. [11] mention just-in-time compute and storage capacity, reducing management and administration cost through automation and providing greater control over end-user service levels.

Virtualization of the high energy physics grid computing clusters has been studied by many researchers. Fenn et al. [12] have tested high performance applications (HPC) in clusters that are made of virtual machines. They found KVM to be usable in non I/O intensive loads. There has been many improvements to the KVM I/O since, and nowadays there is a paravirtualized driver for KVM network and disk, improving I/O performance significantly.

Regola et al. studied the use of virtualization in high performance computing (HPC) [13]. They believed that virtualization and the ability to run heterogeneous environments on the same hardware would make HPC more accessible to a bigger scientific community. They concluded that the I/O performance of full virtualization or para-virtualization is not yet good enough for low latency and high throughput applications such as MPI applications.

Nussbaum et al. [14] made another study on the suitability of virtualization on HPC. They evaluated both KVM and Xen in a cluster of 32 servers with HPC Challenge benchmarks. These studies did not find a clear winner but the authors were able to conclude that the performance of full virtulization is far behind that of paravirtualization. Moreover, running workload among different number of virtual machines did not seem to have an effect. Verma et al. [15], [9] have also studied virtualizaton of HPC applications. They focused on power aware dynamic placement of virtual machines between physical hosts. Though these tests were made with low memory footprint applications, they demonstrated the benefits of virtualization on energy-efficiency. A similar paper by Lui et al. [16] studied the cost of moving virtual machines between physical machines and modelled the energy consumption of virtual machine replacament. The study showed that the cost of migration, meaning the movement of virtual machines between physical hosts, depends mainly on three things; application memory usage, application memory footprint, and network speed.

Padala et al. [17] carried out a performance study of virtualization. They studied the effect of server load and virtual machine count on multi tier application performance. They found OS virtualization to perform much better than paravirtualization. The overhead of paravirtualization is explained by L2 cache misses, which in the case of paravirtualization increased more rapidly when load increased.

Another study from Deshane et al. [18] compares scalability and isolation of a paravirtualized Xen and a full virtualized KVM server. Results said that Xen performs significantly better than KVM both in isolation and CPU performance. Surprisingly, a non-paravirtual system outperforms Xen in I/O performance test.

Virtualization in multi processor environment was studied by Petrides et al. [19]. They found out that virtualization

can be used to stabilize performance for HPC load when executing multiple threads in multi processor environment. In their study the possibility of binding processes and threads to certain cores or processors on the operating system level has not been studied. Nevertheless the study shows how virtualization can improve performance of a multicore and multiuser environment.

As we can see from previous studies, the topic of this paper, the energy-efficiency of virtualization on a single server has not yet received much attention among existing studies.

III. Tests and Test Environment

Our tests aimed at measuring the energy consumption and overhead of virtualization with a diverse test set. We used both synthetic and real applications in our tests and measured how performance is affected by virtualization. We compared the results of the measurements, that were done on virtual machines, with the results of the same tests on physical hardware. We started by measuring the idle consumption of virtualized machines using different number of virtual machines. After that, we compared different virtualization technologies and operating systems.

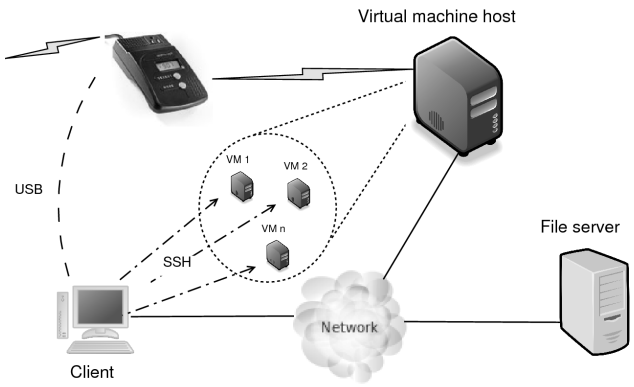


Figure 2. Test environment

For running the tests and collecting measurements, we have used a separate client machine that is connected to the test servers with gigabit local area network. Figure 2 shows the test environment. Cient machine controls how many virtual machines are used and how many applications are run in parallel on the virtual machine host or virtual machines. The client both starts the tests and collects energy meter values. Power usage data was collected with a Watts up? PRO meter via a USB cable. Power usage values were recorded every second.

A. Test Hardware

In our tests we used diverse server hardware. We had both dual CPU servers and single CPU servers. For the Dell 210 II single processor server we had two different types of processors. These two processors represent two different approaches; the E31260L is energy-efficient, while the E31280 is for higher performance. This collection of different types of servers and processors allowed us to study the effect of processor and server architecture more thoroughly.

In our test we used following servers:

- Dell PowerEdge R410, Intel Xeon E5520 w/o Hyperthreading, 16 GB memory, 250 GB hard disk
- 2CPU 12 core server, Opteron 2427, 32GB 800MHz memory, 1TB hard disk
- Dell Poweredge R210, Xeon X3430, 8GB 1333MHz DDR3, 250GB hard disk
- Dell Poweredge R210 II, Xeon E31260L, 8GB 1333MHz DDR3, 1TB hard disk
- Dell Poweredge R210 II, Xeon E31280, 8GB 1333MHz DDR3, 1TB hard disk

B. Used Virtualization Technologies

The operating system used in all machines, virtual or real, was a standard installation of 64-bit Ubuntu Server 10.04.3 LTS. The same virtual machine image was used with both KVM and Xen guests. The image was stored in a raw format, i.e., a plain binary image of the disk image. Linux 3.0.0 kernel was chosen as it had the full Xen hypervisor support. With this kernel we were able to compare Xen with KVM without a possible effect of different kernels on performance.

For CMSSW tests and idle tests, a virtual machine with Scientific Linux 5 was installed with CMSSW version 4.2.4. For these tests real data files produced by the CMS experiment were used. These data files were stored on a Dell PowerEdge T710 server and shared to the virtual machines with a network file system, NFSv4.

C. Test Applications

Our synthetic test collection consisted of Linpack [20], BurnInSSE [1], Bonnie++ [21] and Iperf [22]. Processor performance was measured with an optimized 64-bit Linpack test. This benchmark was run in sets of thirty consecutive runs and power usage was measured for whole sets. In addition, processor power consumption measurements were conducted with ten minute burn-in runs with 64-bit BurnInSSE collection using one, two and four threads. Disk input and output performance were measured using Bonnie++ 1.96. The number of files for a small file creation test was 400. For a large file test the file size was 4 GB. For Bonnie++ tests, the amount of host operating system memory was limited to 2.5 GB with a kernel parameter and

[1] http://www.roylongbottom.org.uk

the amount of guest operating system memory was limited to 2 GB. For hardware tests, a kernel limit of 2 GB was used. The tests were carried out ten times. Network performance was measured using Iperf 2.0.5. Three kinds of tests were run: one where the test computer acted as a server, another where it was the client and a third where the computer did a loopback test within itself. Testing was done using four threads and a ten minute time span. All three types of tests were carried out five times.

As real world applications, we used two different systems. The first one was based on the Invenio document repository [23]. We used an existing Invenio installation, connected to copy of a large bibliogaphic database called Inspire. The Invenio document repository software suite was v0.99.2. The document repository was run on an Apache 2.2.3 web server and MySQL 5.0.77 database management system. All this software were run on Scientific Linux CERN 5.6 inside a chroot environment. Another server was used to send HTTP requests to our test server. The requests were based on an anonymous version of a real-life log data of the identical document repository in use at CERN. The requests were sent using the Httperf web server performance test application [24].

Table I shows the rates and resources given to virtual machines in the Invenio tests. The MaxClients setting refers to the maximum clients setting in Apache web server configuration.

Table I
SETTINGS FOR CHANGING LOAD AND RESOURCES OF A SINGLE VIRTUAL MACHINE

VCPUs	Memory (GB)	MaxClients	Request rate
2	5	8	5
4	8	15	10
6	15	24	15

The second real application was a physics data analysis that used the CMSSW framework [25]. This analysis task is a typical one in high-energy physics. We used real data created at CERN. The data was stored in a ROOT image[26] files, which our case were of size 4GB. Normally, a data analysis with this data takes days to perform, thus we limited the number of events of one analysis task to 300. With this limitation the analysis takes 10 minutes on the Opteron hardware. The data was located on network file system, NFS, and reading it caused very little network traffic, 2kB per task.

Tables II and III show how the resources of the 12-core Opteron server were shared between virtual machines when testing the constant load with different number of virtual machines running the CMSSW tests. In all the cases the hardware resources were shared equivalently among different virtual machines.

Table II
SETTINGS FOR A SINGLE VIRTUAL MACHINE IN 12-CORE OPTERON SERVER

VM count	VCPUs	Memory (GB)
1	12	31.5
4	3	7.88
8	2	3.94
10	2	3.15
12	1	2.6

Table III
SETTINGS FOR A SINGLE VIRTUAL MACHINE IN 4-CORE DELL 210 II

VM count	VCPUs	Memory (GB)
1	8	7.5
2	4	3.75
3	3	2.50
4	2	1.88
5	2	1.50
6	1	1.25

IV. RESULTS

A. Idle consumption

First we studied idle energy consumption with different virtualization solutions and with different number of virtual machines. Figure 3 shows the power consumption of two different virtualization solutions, and the power consumption of the hardware with no virtual machines. In this test both the host system and the virtual machines were running only basic operating system functions without any analysis tasks. In all the measurements with virtual machines we had three virtual machines running idle. The figure shows how energy consumption of two different virtualization solutions behave when the servers are idle. It shows how overhead of virtualization depends on the virtualization solution and kernel version. The difference between KVM and hardware is less than 3%, which is already a big improvement compared to three separate physical machines running idle. This test was run with the Dell R410 server.

The second idle measurement was run on the dual processor Opteron server. The test measured the energy consumption of an idle physical hardware for 20 minutes. Figure 4 shows how the operating system affects the idle consumption. The same test was repeated with different number of virtual machines on the same physical hardware. The energy consumption accumulates with the virtual machine count with Scientific Linux 5 (SLC5) but with Ubuntu it remains almost the same as with bare hardware. This test shows that the choice of virtual machine has big effect on energy-efficiency. It also shows that an idle virtual machine can have a very low energy consumption.

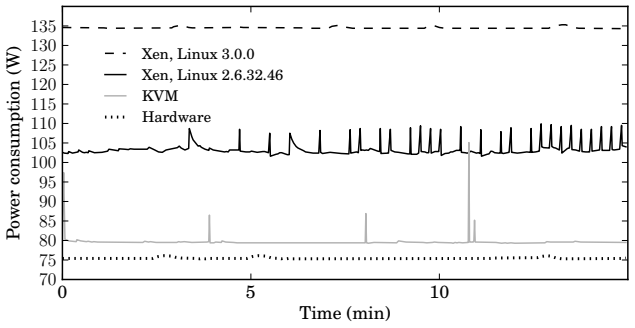


Figure 3. Typical idle power consumption

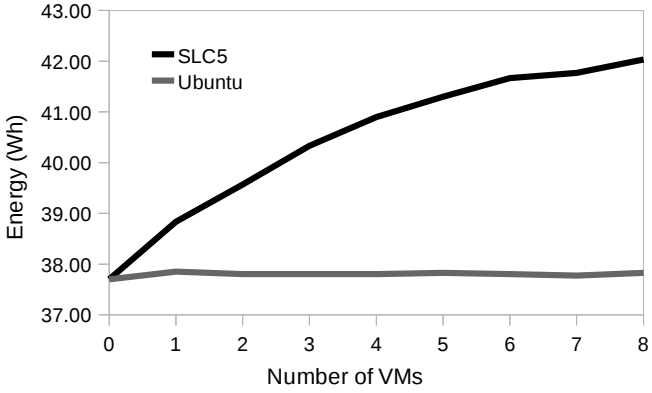


Figure 4. Energy consumption of idle virtual machines

B. Synthetic tests

We used synthetic tests to stress different server components; CPU, I/O and network. With these tests we studied in which situations virtualization causes the most overhead. First we tested the overhead of disk reads and writes with Bonnie++. The consumption of the virtualization solutions; Xen and KVM was compared to hardware consumption.

In images from 5 to 12 the y-axis represents the percentual difference to corresponding measurents done without virtualization.

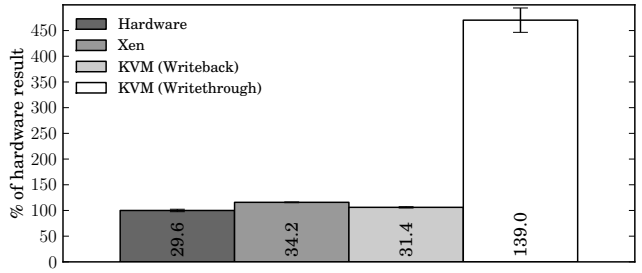


Figure 5. Energy consumption of Bonnie++ (Wh)

As can be seen from Figure 5, when running a set of synthetic disk operations Xen uses slightly more energy compared to hardware. With KVM the situation is different. When using the default cache setting, write through cache, KVM uses around 350% more energy than hardware. About 90% of the test time is spent doing the small file test part of Bonnie++. Switching to write back cache, results of KVM are actually slightly better than hardware results. Write back cache writes only to a cache and stores data to the disk only just before the cache is replaced. This is a cache mode that is not safe for production use and is available mainly for testing purposes.

Next, we tested the overhead of virtualization of CPU with two different benchmarks; BurninSSE and Linpack. Figure 6 shows the power consumption of the server while running BurninSSE on a non-virtualized server and on virtual machines. We compared the technologies by introducing CPU load with 1 and 4 threads of BurnInSSE. With 1 thread, KVM and hardware use the same amount of power, but Xen uses around 10% more. With 4 threads the situation is the other way around: Xen uses less power than KVM and hardware. The explanation can be seen in Figure 7. Even though Xen uses more power in the single-threaded LINPACK test, it is slower: the CPU is not running at its full turbo boosted speed, but Xen has a systematic overhead in power consumption compared to the others. With 4 threads, Xen's CPUs are not running at full speed so the power usage is not as great as with hardware or KVM, and the overhead in power consumption is overshadowed by the power usage of 4 computing threads.

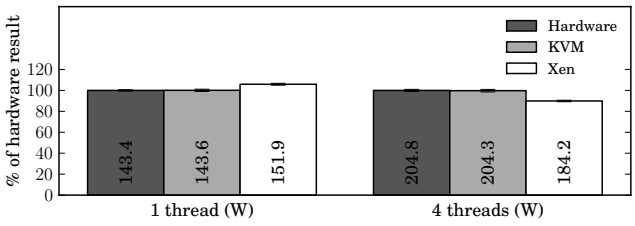


Figure 6. Power consumption under high CPU load with BurnInSSE

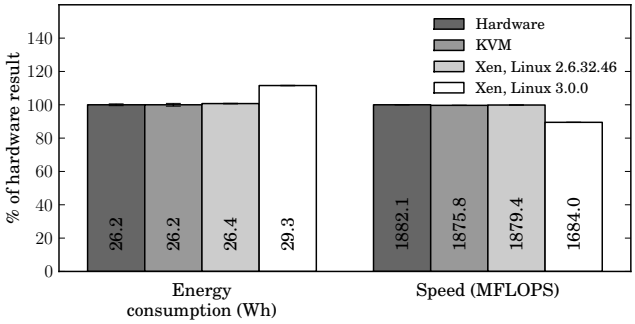


Figure 7. Energy consumption under high CPU load with Linpack

As the last server component we tested the network and

the overhead of virtualization to the power consumption of the network. To stress the network we used Iperf network benchmark in dual mode, where the traffic is tested to both directions. The power consumption of Iperf test results are shown in Figure 8. Direction of the traffic does not have an effect on the results, which show a similar trend for I/O and CPU tests: KVM uses slightly more power than hardware while Xen consequently uses slightly more power than KVM. Interestingly, when a Xen virtual machine was running as server it used slightly more power than when running as client. With KVM and hardware it was the other way around. In the loopback mode, one can find similar results with Xen as in the LINPACK test in Figure 7: for some reason, Xen's performance is capped and consequently bandwidth in the loopback mode is much worse than with KVM or hardware, and on the other hand mean power consumption is lower.

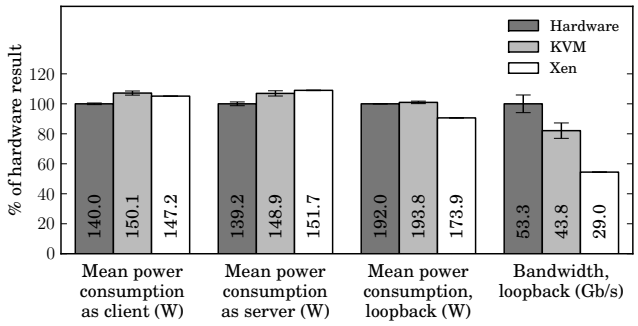


Figure 8. Power consumption under Iperf network traffic test

C. Realistic load

Realistic tests were designed such that we would get better understanding of energy usage in two different real world situations: web services and physics analysis. Both benefit from virtualization differently as they use server resources in a different way. Web server based systems benefit from being able to combine idle services into single physical server and the overhead of virtualization is not so critical. The physics analysis benefits from an isolated and job specific environment provided by virtualization, but this causes it to run a longer time.

We started our realistic tests with the Invenio CERN document server repository case. In this test, we sent HTTP requests, which were based on CERN library log data, to a virtualized web server. We measured both performance and power consumption. We ran the same tests with and without virtualization. We compared two virtualization solutions to hardware to measure the overhead of virtualization. In all the Invenio tests, the Invenio installation was in a chroot environment with a complete SLC5 installation. To assure that chroot between the operating system and the Invenio web application did not have any negative effects on test results, a comparative test was performed between the base system and another chroot environment using a copy of the base system as the new root.

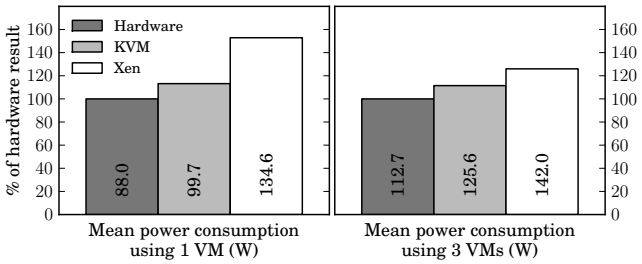


Figure 9. Power consumption of different virtualization solutions with different number of virtual machines in the repository test

Both the amount of virtual machines and virtualization technology have an impact on the energy-efficiency. This effect was tested here by running Invenio document server in different number of virtual machines and with different virtualization technologies. Figure 9 shows how the power consumption varies between different virtualization technologies. On the left side we have the results of running one virtual machine with a rate of 5 queries per second workload and on the right side 3 virtual machines with a rate 5 queries per second per virtual machine and total request rate of 15 queries per second. Figure shows that the power consumption evens out when we have more virtual machines and load.

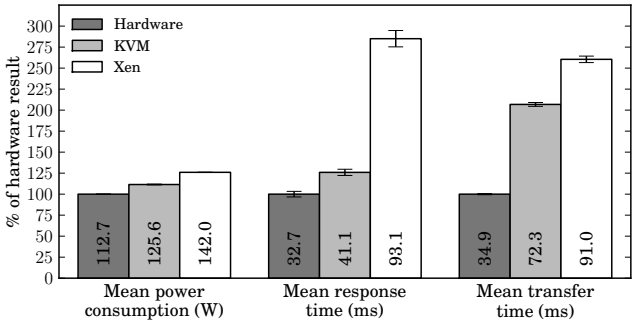


Figure 10. Power consumption and Httperf results of different virtualization solutions

Figure 10 illustrates the effect of virtualization on the web performance when running Invenio on three virtual machines with request rate of 5 on each. In the figure we have both the energy consumption from the previous figure and the results from the Httperf test. One can see that even though there is not much overhead on energy consumption there is a bigger impact to the quality of service as the response times and transfer times increase.

Previous results showed us that KVM performed closer to hardware level and proved to be more interesting for

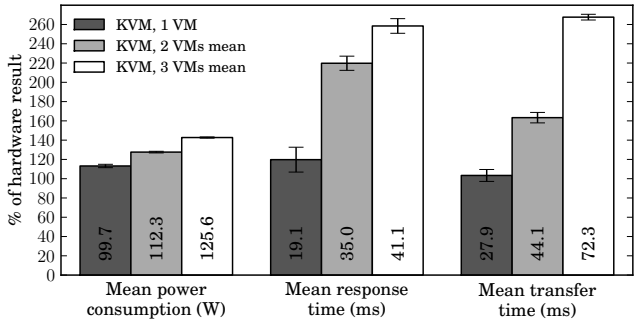


Figure 11. The effect of workload on virtual machine performance with KVM using different amounts of virtual machines

further studies. KVM was used in our test where we studied the overhead of virtualization by increasing the number of virtual machines with similar workload. These results are illustrated in Figure 11, showing that the power consumption increased linearly as a function of virtual machines. In the case of 3 virtual machines the total consumption decreases of 47%, but on the other hand the response times and transfer times increase more than 250%.

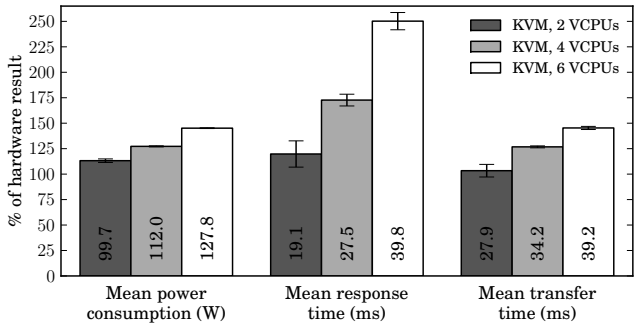


Figure 12. The effect of workload on virtual machine performance with KVM using different virtual machine resources

Virtual machines share server resources and the share of resources given to a virtual machine is configurable. One can either stretch the physical resources thin between several virtual machines or one can create a few virtual machines with more resources. Table I shows how the resources were shared. Figure 12 illustrates how the performance and energy-efficiency is affected when we increase the resources and load of a single virtual machine. Here you can see that the power consumption grows almost the same way as in the Figure 11. Difference being that increasing the resources of one virtual machine seems to improve the performance.

To illustrate the effect of virtualization on quality of service we have Figures 13 and 14. These figures show a cumulative distribution of response times that the Httperf test application reported for the HTTP requests. The distribution shows how the response times behave with different virtualization technology and load.

In Figure 13, we have the distribution of response times from a test with 3 virtual machines and total request rate of 15 requests per second. This distribution corresponds to the results illustrated by the right side of Figure 9 and Figure 10. One can see that both KVM and Xen decrease quality of service, but still more than 95% of requests are served in 100ms.

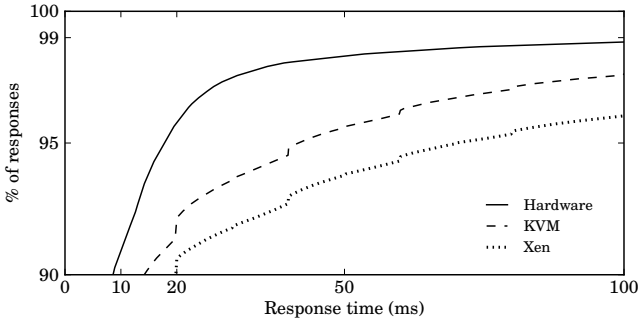


Figure 13. The impact of virtualization solution on quality of service

To show how virtual machine count affect the quality of service, we made a similar distribution from the test that was illustrated by Figure 11. Figure 14 shoes how the virtualization overhead effect on different workload and number of virtual machines. We compared KVM and one to three virtual machines with corresponding rates 5, 10, and 15 requests per second. Every additional virtual machine decreased the quality of service.

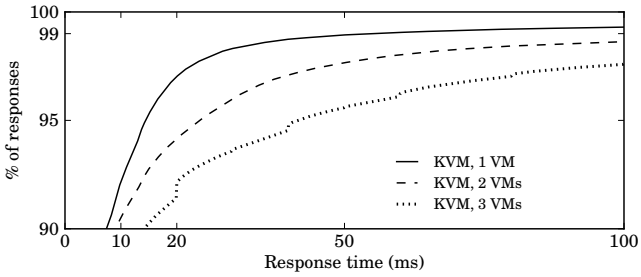


Figure 14. The impact of different loads on quality of service

As our second realistic load, we had a physics analysis application, CMSSW. First we tested how the number of virtual machines affects the performance of CMSSW. In the following tests we consider one run of the test application as a job. In Figure 15, we have the results of running 15 jobs in 5 different virtual machine sets and also on hardware. The figure shows how the energy efficiency degrades as the number of virtual machines increases and, at the same time, throughput decreases. 15 smaller virtual machines running one job are 6.8 times less energy efficient than running 15 jobs on one bigger virtual machine.

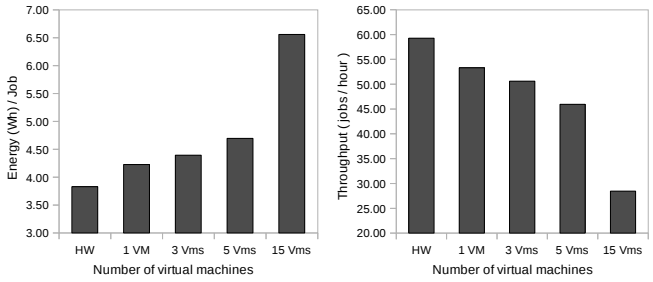


Figure 15. Running 15 jobs in different number of virtual machines

Next, we tested the effect of dual processor architecture on the overall performance. We run the same tests both on 2 CPU 12 core Opteron server and on single processor quad core R210 server. This test differs from the previous not only by the hardware, but also by the load introduced. Here we started with a very low load, that was increased to see how the overhead behaves on lower load. Figures 16 and 17 show the energy consumption and throughput from a test with different amount virtual machines running one job each. One can see that the virtualization introduces some overhead on both servers and this overhead increases as the amount virtual machines is increased. Single CPU server's performance is limited by the 8GB memory as a single CMSSW analysis job together with the operating system use approximately 1.2GB of memory. Figure 16 shows the results from both virtualized servers and physical hardware.

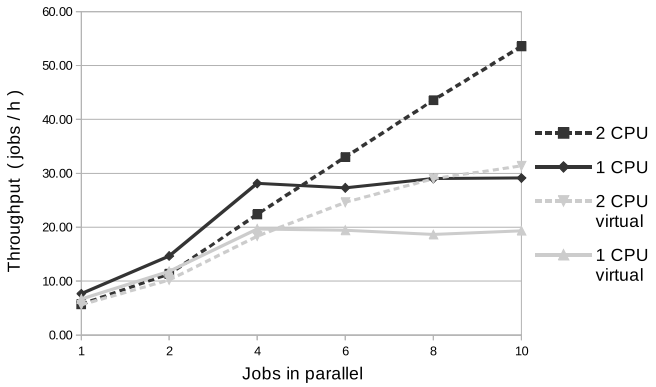


Figure 16. Throughput with different number of virtual machines running one job each

As shown in Figure 17, the single processor server has a better energy-efficiency on low loads, but this balances when the load is increased. One thing to notice is how static the overhead is on single processor server. In two processor server overhead increases as a function of virtual machines.

In the previous tests, we used light load on virtual machines, but used them in high numbers. Now we show how physics analysis job behaved in different sized virtual machines. As the Invenio tests showed, bigger virtual machines

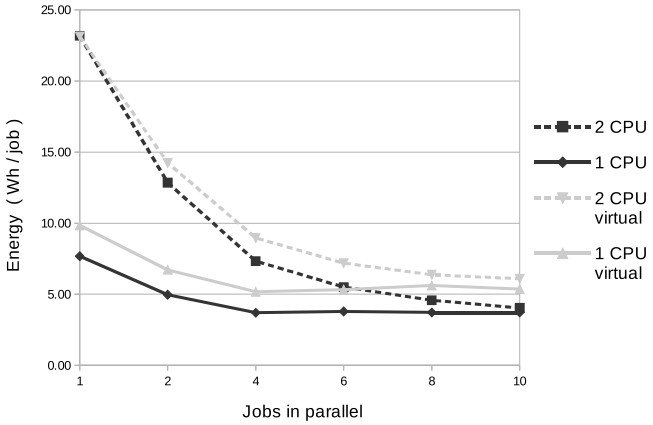


Figure 17. Energy consumption per job with different number of virtual machines running one job each

perform better. Figure 18 shows the effect of workload on energy-efficiency with physics analysis software. We tested different workloads on 5 identical virtual machines sharing the Opteron 12 core server. One can see that the energy-efficiency and throughput improve as we increase the load. This is in line with our earlier studies where we noticed that the commonly used one job per CPU core does not give the best performance or energy efficiency [27], [28]. Here we tested how it applies to virtualized environments.

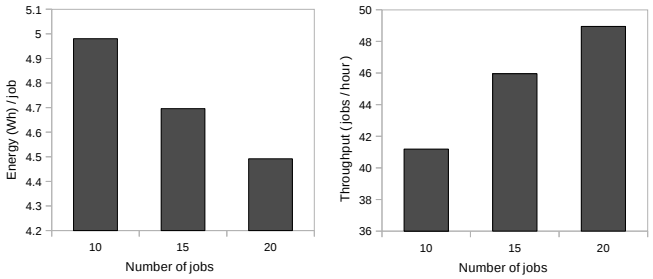


Figure 18. Running different workload on 5 virtual machines

As the number of virtual machines and their size seem to make a difference on dual processor server we repeated the test on single processor server. Figures 19 and 20 illustrate the effect of parallelism with rising load. Here we have run different amount of jobs on one physical server and divided the load equally between one, two, four and six virtual machines. One can see that the performance increases when the load is increased up to a point where the system throughput levels and eventually decreases. One can notice that the amount of virtual machines has big effect on the throughput and energy-consumption. The virtualization overhead increases exponentially as the function of virtual machine count and increases even more when the number of virtual machines is more than the number of cores in the server's processor.

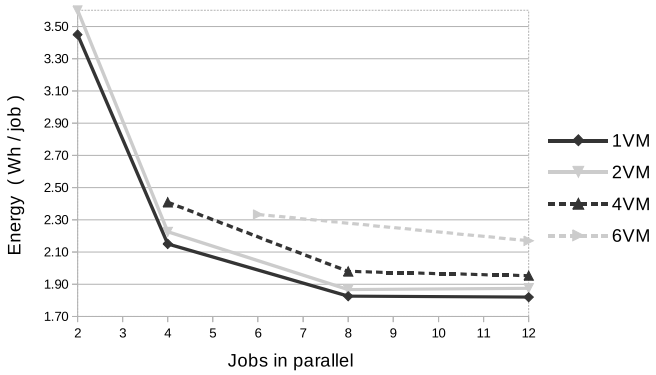


Figure 19. Energy consumption per job in virtual machines when varying resources and job parallelism

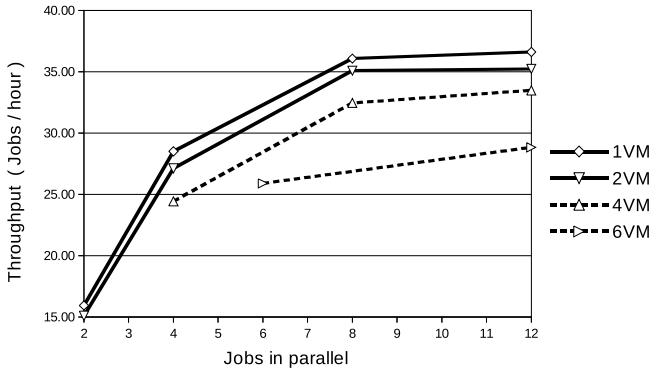


Figure 20. Job throughput of virtual machines with varying resources and job parallelism

These single processor tests above were performed with two types of CPUs; energy-efficient and powerful ones. Results shown in Figures 19 and 20 are from tests run with energy-efficient processor. The results from the powerful processor were similar to those of the energy efficient processor and produced similar figures. The difference was the energy-efficient processor used 17% less energy per job and the throughput of the powerful processor 40% better when comparing the minimum of energy consumption and the maximum of throughput.

We ran an additional test in parallel with the physics tests to study the reason behind the overhead. We tested how the network performance is affected by the increasing load and increasing number of virtual machines. This was done by running a ping command from the virtual machines towards the client machine. This was tested by running 12 jobs equally among different amount of virtual machines; 1,4,6 and 12. We noticed that the latencies did not drop while adding more virtual machines. This test showed that the resource sharing between virtual machines was fair and latencies varied very little between virtual machines.

V. Conclusions

Virtualization technologies develop at fast pace. New technologies arise and better interfaces are made to improve the usability of virtualization. In this study we have used two mature open source virtualization solutions; KVM and Xen. The performance of Xen and paravirtualization have been good for a long time, but for the version used in our tests suffered from the early adoption on Linux kernel and had not had enough time to mature in the vanilla Linux kernel. KVM on the other hand have come far from its early versions and proves comparable with the commercial virtualization solutions. Even though the technologies in this study were compared and tested against each other this should not be considered as a comparison between different virtualization technologies, but as a study on virtualization technologies in general. The performance balance between different technologies varies constantly, but the main idea is that resources are shared among multiple systems and this causes overhead to applications inside virtualized servers, which needs be taken into consideration.

The overhead of virtualization is a well-known fact and reported in many publications. Although the technologies have been improving a lot during the past five years, the performance of a virtualized system is still far from the hardware level. However, this does not mean that virtualization could not be useful in improving energy-efficiency in large data centers but it means that one should know how to apply this technology to achieve savings in energy consumption.

We studied the energy-efficiency of virtualization technologies and how different loads affect it. Our research indicates that idle power consumption of a virtualized server is close to zero. However, this depends a lot on the operating system running on the virtual machine, but it is always a small number compared to idle energy consumption of a physical server. Our study also indicates that virtualization overhead has great impact on energy-efficiency. This means that it would make more sense to share the physical resources among few virtual machines with heavy load instead of a larger set of light-loaded ones. Pure CPU-load in larger virtual machine groups does not seem to impose as much overhead as the more complex physics analysis job, that both requires network connectivity and disk storage. The physical core count also seem to pose a limit for the virtual machine pool size.

Acknowledgment

Many thanks to Jochen Ott from CMS@CERN experiment for providing help in installing the CMSSW and providing with a typical analysis job. Also we would like to thank Salvatore Mele, Tibor Simko, Jean-Yves LeMeur of CERN library and Invenio developers, for providing realistic data and a test case for our analysis; and Marko Niinimaki for comments.

REFERENCES

[1] J. Kommeri, T. Niemi, and O. Helin, "Study of virtualization energy-efficiency in high energy physics computing," in *Proc. Energy 2012*, 2012.

[2] J. Kommeri, T. Niemi, and M. Niinimki, "Study of virtualization energy-efficiency in high energy physics computing," in *Proc. ICGREEN'12*, 2012.

[3] B. Schäppi, F. Bellosa, B. Przywara, T. Bogner, S. Weeren, and A. Anglade, "Energy efficient servers in europe," Austrian Energy Agency, Tech. Rep. October, 2007.

[4] E. STAR, "Report to congress on server and data center energy efficiency," U.S. Environmental Protection Agency ENERGY STAR Program, Tech. Rep., 2007.

[5] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *Computer*, vol. 40, pp. 33–37, 2007.

[6] W. Vogels, "Beyond server consolidation," *Queue*, vol. 6, pp. 20–26, January 2008.

[7] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," in *Proceeding of the 14th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS '09. Washington, DC, USA: ACM, 2009, pp. 205–216.

[8] U. Hölzle and B. Weihl, "High-efficiency power supplies for home computers and servers," Google, Tech. Rep., 2006.

[9] A. Verma, P. Ahuja, and A. Neogi, "pmapper: power and migration cost aware application placement in virtualized systems," in *Proceedings of the 9th ACM/IFIP/USENIX International Conference on Middleware*, ser. Middleware '08. New York, NY, USA: Springer-Verlag New York, Inc., 2008, pp. 243–264.

[10] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering it services as computing utilities," in *High Performance Computing and Communications, 2008. HPCC '08. 10th IEEE International Conference on*, sept. 2008, pp. 5 –13.

[11] M. Xu, Z. Hu, W. Long, and W. Liu, "Service virtualization: Infrastructure and applications," in *The grid: blueprint for a new computing infrastructure*. Wiley, 2004, ch. 14.

[12] M. Fenn, M. A. Murphy, and S. Goasguen, "A study of a kvm-based cluster for grid computing," in *Proceedings of the 47th Annual Southeast Regional Conference*, ser. ACM-SE 47. New York, NY, USA: ACM, 2009, pp. 34:1–34:6. [Online]. Available: http://doi.acm.org/10.1145/1566445.1566492

[13] N. Regola and J.-C. Ducom, "Recommendations for virtualization technologies in high performance computing," in *Cloud Computing Technology and Science (CloudCom), 2010 IEEE Second International Conference on*, 30 2010-dec. 3 2010, pp. 409–416.

[14] L. Nussbaum, F. Anhalt, O. Mornard, and J.-P. Gelas, "Linux-based virtualization for hpc clusters," *Network*, pp. 221–234, 2009.

[15] A. Verma, P. Ahuja, and A. Neogi, "Power-aware dynamic placement of hpc applications," in *Proceedings of the 22nd annual international conference on Supercomputing*, ser. ICS '08. New York, NY, USA: ACM, 2008, pp. 175–184.

[16] H. Liu, C.-Z. Xu, H. Jin, J. Gong, and X. Liao, "Performance and energy modeling for live migration of virtual machines," in *Proceedings of the 20th international symposium on High performance distributed computing*, ser. HPDC '11. ACM, 2011, pp. 171–182.

[17] P. Padala, X. Zhu, Z. Wang, S. Singhal, and G. Shin, K., "Performance evaluation of virtualization technologies for server consolidation," *Work*, no. HPL-2007-59, p. 15, 2007.

[18] T. Deshane, Z. Shepherd, J. Matthews, M. Ben-Yehuda, A. Shah, and B. Rao, "Quantitative comparison of xen and kvm," in *Xen summit*. USENIX association, June 2008.

[19] P. Petrides, G. Nicolaides, and P. Trancoso, "Hpc performance domains on multi-core processors with virtualization," in *Proceedings of the 25th international conference on Architecture of Computing Systems*, ser. ARCS'12, 2012, pp. 123–134.

[20] J. Dongarra, P. Luszczek, and A. Petitet, "The linpack benchmark: past, present and future," *Concurrency and Computation Practice and Experience*, vol. 15, no. 9, pp. 803–820, 2003. [Online]. Available: http://doi.wiley.com/10.1002/cpe.728

[21] B. Martin, "Using bonnie++ for filesystem performance benchmarking," *Linuxcom*, vol. Online edi, 2008.

[22] M. Egli and D. Gugelmann, "Iperf - network stress tool," *Source*, pp. 1–2, 2007.

[23] J. Caffaro and S. Kaplun, "Invenio: A modern digital library for grey literature," in *Twelfth International Conference on Grey Literature*, Prague, Czech Republic, Dec 2010.

[24] D. Mosberger and T. Jin, "httperf - a tool for measuring web server performance," *SIGMETRICS Perform. Eval. Rev.*, vol. 26, pp. 31–37, Dec 1998.

[25] F. Fabozzi, C. Jones, B. Hegner, and L. Lista, "Physics analysis tools for the cms experiment at lhc," *Nuclear Science, IEEE Transactions on*, vol. 55, pp. 3539–3543, 2008.

[26] I. Antcheva and et al., "Root a c++ framework for petabyte data storage, statistical analysis and visualization," *Computer Physics Communications*, vol. 180, no. 12, pp. 2499 – 2512, 2009.

[27] T. Niemi, J. Kommeri, K. Happonen, J. Klem, and A.-P. Hameri, "Improving energy-efficiency of grid computing clusters," in *Advances in Grid and Pervasive Computing, 4th International Conference, GPC 2009, Geneva, Switzerland*, 2009, pp. 110–118.

[28] T. Niemi, J. Kommeri, and H. Ari-Pekka, "Energy-efficient scheduling of grid computing clusters," in *Proceedings of the 17th Annual International Conference on Advanced Computing and Communications (ADCOM 2009), Bengaluru, India*, 2009.

Energy and Carbon Aware Scheduling in Supercomputing

Mikko Majanen, Olli Mämmelä
Autonomic Networking Team
VTT Technical Research Centre of Finland
Kaitoväylä 1, Oulu, Finland
Email: mikko.majanen@vtt.fi, olli.mammela@vtt.fi

André Giesler
Jülich Supercomputing Centre
Forschungszentrum Jülich
52425 Jülich, Germany
Email: a.giesler@fz-juelich.de

***Abstract*—At an early stage of information and communications technology and high-performance computing, performance and reliability were two important factors in research and development. Energy consumption was not considered as a serious topic, since the technical characteristics of hardware and software were limited and the amount of computing nodes in a computing cluster, i.e., a data centre was small. Gradually the situation has evolved a lot: nowadays there are multiple data centres located in geographically diverse locations and the software has become more complex. Modern data centres are equipped with a large amount of computing nodes having vast computing power. Consequently, energy consumption has become a major topic. This work presents two algorithms for optimizing energy and emissions in high-performance grid computing, in which multiple data centres are interconnected to each other. The algorithms are validated both in simulation and testbed environments. The effect of various parameters to energy and emission savings are studied and the performance of the algorithms is compared to commonly used default algorithms. Our simulation and testbed experiments show that the developed algorithms are able to reduce energy consumption and emissions drastically without significant increase in job turnaround or wait time.**

Keywords-HPC; grid computing; energy; emissions; testbed.

I. Introduction

The increased demand for IT applications and services has encouraged the building of data centres worldwide. However, data centres consume an enormous amount of energy at an increasing financial and environmental cost. This has led to research efforts in both industry and academia to cut down data centre energy usage and emissions. This work is a continuation to our prior work in ENERGY 2012 [1] to save energy and reduce the CO_2 emissions in federated High-Performance Computing (HPC) data centres. Previously, we have also researched the energy saving potential inside single site data centres [2].

In 2006, U.S. servers and data centres consumed around 61 billion kilowatt hours (kWh) at a cost of about 4.5 billion U.S. Dollars [3]. This is equal to about 1.5% of the total U.S. electricity consumption or the output of about 15 typical power plants. High energy consumption naturally causes huge environment pollution. It has been estimated that Information and Communications Technology (ICT), as a whole, covers 2% of world's CO_2 emissions [4] and this amount looks set to grow at 6% each year until 2020 [5]. Data centres were 14% of the total ICT footprint in 2002 and 2007, and it is estimated that the amount will rise to 18% in 2020.

In HPC, the ever-growing demand for higher performance seems to increase the total power consumption, even though more flops per watt are achieved. In order to provide even greater computing capabilities, HPC data centres can be interconnected to each other to form larger, federated or HPC grid data centres. The connection is implemented by using special grid software (e.g., UNICORE (UNiform Interface to COmputing REsources) [6]) that manages the job submissions to all data centres belonging to the grid.

The energy consumption between the data centres may vary radically due to the different characteristics of the centres. For example, the server hardware in each centre may be different and consume different amount(s) of energy. The centres may also locate geographically far from each other and the surrounding climate can cause large differences in the needed cooling, i.e., the Power Usage Effectiveness (PUE) [7] values between different centres may vary due to the surrounding climate. Also, since the energy sources can differ between the centres, the CO_2 emissions of the data centres may vary radically depending on the available energy sources. The differences between the data centres naturally enable optimizations regarding energy consumption and CO_2 emissions.

In our prior work [1] we introduced two algorithms for selecting the data centre inside the grid in energy- and CO_2-aware manner. The performance of the algorithms was studied by simulations and the results showed significant savings in energy consumption and CO_2 emissions. This work extends our previous work by a feasibility study [8] of the algorithms in a testbed environment that consists of three clusters: two located in Germany and one in Finland. The testbed results confirm the

possibilities for energy and emission savings achieved in the simulations, and can be used as a basis for the design of specific federated cluster environments when using a developed software plug-in to enable an energy-aware scheduling of the resources. Furthermore, we also consider energy and CO_2 emission savings inside single HPC data centres by using energy-aware scheduling algorithms presented in [2].

The rest of the paper is organized as follows: Section II describes the related work. Section III introduces the cluster selection algorithms. Section IV describes the simulation model and scenario, and presents the simulation results. Testbed experiments are presented in Section V, including the scenario and results. Conclusion and future work are presented in Section VI.

II. Related work

As described in [2], several methods for saving energy in single HPC data centres have been studied. The methods include mainly the use of energy-efficient or energy proportional hardware (e.g., embedded low-power chips), Dynamic Voltage and Frequency Scaling (DVFS) techniques, shutting down idle hardware components at low system utilizations, power capping, and thermal management. Recently, there has also been approaches to solve HPC energy issues in Graphics Processing Unit (GPU) computing [9], [10], [11]. GPU computing aims at combining the use of a GPU with a CPU to accelerate general-purpose scientific and engineering applications. The compute-intensive portions of the application are offloaded to the GPU, while the remainder of the code still runs on the CPU. However, the energy consumption is a major concern in these systems.

In our prior work [2], we used an energy-aware job scheduler to schedule the jobs inside single data centres and shut down idle computing nodes whenever possible. We also noted that merely the choice of a different scheduling algorithm can affect the energy consumption of a data centre. Out of commercial HPC schedulers, Moab offers a Green Computing plug-in [12] that tries to reduce power consumption and costs in a data centre in a quite similar way as our energy-aware job scheduler. In the Moab plug-in, it is possible to turn off idle nodes that do not have reservations on them, and turn on additional nodes when jobs require them. Moab uses a MAXGREENSTANDBYPOOLSIZE parameter, where users can specify a "green pool", which is the number of nodes that are kept on and ready to run jobs (even if some nodes are idle). Idle nodes that exceed the number specified with the MAXGREENSTANDBYPOOLSIZE parameter are turned off. The requirements for the Green Computing plug-in are a license for green computing, Moab 5.3.5 or later, a script that Moab can call to programatically turn nodes on and off, and a resource manager that can monitor and report power state. In a test run, the Green Computing plug-in was able to decrease the energy consumption by 8.2% with the penalty of 7.5% increased workload completion time [13]. The savings with Moab Green Computing depend highly on the workload.

In this paper we extend our scope from single HPC data centres to HPC grid data centres and introduce two algorithms for selecting the data centre inside the grid in energy- and CO_2-aware manner. Moreover, we provide HPC grid simulation and testbed results, and new single HPC data centre results.

Until recently, there has not been much previous research that addresses the energy efficiency or CO_2 emissions of the grids from the whole grid perspective; mainly only optimizations inside a single data centre have been studied. Perhaps the most similar approach to our approach is the Heterogeneity Aware Meta-scheduling Algorithm (HAMA) [14]. HAMA first selects the most energy-efficient cluster for the job based on the power consumption of the servers and the efficiency of the cooling system. Additionally, when running the job, DVFS is used to reduce the power consumption of the CPU. The simulation results show that HAMA can reduce up to 23% energy consumption in the worst case and up to 50% in the best case as compared to other algorithms (EDF-FQ, which prioritizes jobs based on a deadline and submits jobs to resource sites in earliest start time (FQ) manner with the smallest waiting time). Without DVFS, HAMA can still result in power savings of up to 21%.

Lynar *et al.* [15] have explored the effect on energy consumption by using different resource allocation mechanisms, both in a cluster and in a grid. The results show that different resource allocation methods can result in a significantly different energy usage while computing a stream of tasks. The Pre-processed Batch Auction (PPBA) and batch auctions almost always result in a significantly lower energy use than a random resource allocation. By using a simple batch auction allocation method, energy consumption can be reduced by up to 37.5%, and possibly even more by using the PPBA method.

Patel *et al.* [16] have presented an energy-aware policy for distributing computational workload in the Grid resource management architecture. They introduce a data centre energy coefficient that is taken into account as a policy when making allocation decisions for compute workloads. This coefficient is determined by the thermal properties of each data centre's cooling infrastructure including regional and seasonal variations. The estimated energy savings in case of three data centres located in two different time zones were large enough to give sufficient reason for the economic viability of the approach.

Shah and Krishnan [17] also analyze the climatic conditions as a means to reducing cooling energy costs. They show that dynamic optimization of the thermal workloads based on local weather patterns can reduce the environmental burden by up to 30% in their case study. Additionally, the data centre operational costs can be potentially reduced by nearly 35%. Due to the variability of fuel mixes encountered in a global grid, they also found that the use of pure energy consumption as a metric for environmental sustainability — a common practice in the ICT literature — can be erroneous.

The GREEN-NET framework [18] consists of an ON/OFF model, which includes prediction heuristics and green advice for the users and takes the decision to switch on or off the nodes, and an adapted energy efficient Resource Management System (RMS) at the grid level.

There has also been research on the feasibility of powering data centres more by renewable energy ([19], [20], [21]) and studying the environmental potential of Geographical Load Balancing (GLB) [22], in which processes are shifted to data centres located in regions where energy currently has low cost. In [23], a renewable and cooling aware workload management plan is introduced. The availability of renewable energy and IT demand is predicted and IT resources are allocated according to a time varying power supply and cooling efficiency. A similar approach is taken in GreenSlot [24], which is a parallel batch job scheduler powered by solar power and the electrical grid (as a backup). It predicts the amount of solar energy that will be available in the near future and schedules the workload to maximize the use of green energy without breaking any Service Level Agreement (SLA).

III. Optimization in the HPC Grid

The optimization algorithm in the HPC grid focuses on optimizing the scheduling process in the UNICORE middleware [6]. The scheduling process is triggered by submitting a job from the UNICORE Commandline Client, or from the UNICORE Rich Client to the UNICORE Workflow Engine. The UNICORE Workflow Engine queries a UNICORE Service Orchestrator (USO), on which cluster the job should be submitted. As a default, the USO uses round-robin algorithm for choosing the cluster. After cluster decision, the job is submitted to the RMS of the chosen cluster. The RMS takes care of executing the job according to the used scheduling algorithm, e.g., FIFO (First In, First Out) or backfilling.

In this work, we focus on reducing the energy consumption and the CO_2 emissions. The CO_2/energy related optimizations should not affect the current SLA or Quality of Service (QoS) agreements, or alternatively, a new green SLA [25] could be used. In HPC, there are no clear SLAs between users and data centres, but a reasonable turnaround time can be seen as sort of a QoS agreement. A possible green SLA for HPC data centres could mean that the users allow certain delay for the execution of their job. As a bonus, they will get some extra computing time for free.

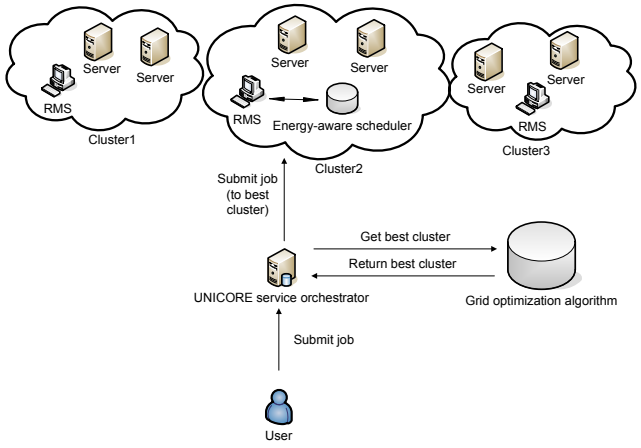


Figure 1. Job submission in a federated HPC data centre

For decreasing CO_2 emissions and/or energy consumption in federated HPC data centres, the optimization algorithm will be used for performing the cluster selection in CO_2/energy-aware manner. In addition to cluster selection algorithms, also energy-aware single site scheduling algorithms will be used. As depicted in Figure 1, the USO in UNICORE receives job requests coming from the users. The jobs include the requirements for the needed resources (e.g., number of nodes/cores, memory, etc.). If the user wants to use the green SLA, it is also included in the job requirements. The grid optimization algorithm is used to select the most suitable cluster for the job and the job is subsequently submitted to the RMS of the selected cluster. The RMS uses energy-aware job scheduling algorithms to schedule the job and power off idle servers. The energy-aware job scheduling algorithms for single site data centres were defined in our previous work [2]. The main principle of the single site algorithms is to keep the system active with the lowest amount of resources as possible. If a node is not needed for job execution, it is shut down or placed into an energy saving mode. Once the node is needed for the job execution again, it is woken up.

A. Carbon Usage Effectiveness

Carbon Usage Effectiveness (CUE) is a sustainability metric developed by the Green Grid organization [26]. The main purpose of the metric is to address carbon emissions associated with data centres. The CUE can be calculated as follows:

Table I
ENERGY EMISSION COEFFICIENT FACTORS

Generation type	Conversion factor ($kgCO_2$ per kWh)
Closed cycle gas turbine	0.360
Coal	0.910
Electricity, France interconnector	0.083
Electricity, Ireland interconnector	0.699
Non pumped storage hydro	0.0
Nuclear	0.0161
Open cycle gas turbine	0.479
Oil	0.610
Pump storage	0.0
Other	0.610

$$CUE = \frac{SiteEmissions}{ICTEnergy}, \quad (1)$$

where $ICTEnergy$ is the energy consumpted by the ICT equipment in the data centre. An alternative approach for calculating the CUE is to multiply the Energy Source Coefficient (ESC) by the data centre's PUE:

$$CUE = ESC * PUE, \quad (2)$$

where PUE is a metric for defining how efficiently the power in the data centre is used, i.e., how much power is actually used by the ICT equipment and how much power is used for cooling and other equipment. ESC is defined as follows:

$$ESC = \sum ESP * EEC, \quad (3)$$

where Energy Source Percent (ESP) indicates the percentage of the energy generation source, and Energy Emission Coefficient (EEC) indicates how many kilograms of CO_2 are emitted per 1 kWh of energy. Example values of the EEC can be found in Table I [27]. By using the formulas described earlier and the values in Table I, we are able to estimate how much emissions are caused by data centres with different energy sources:

$$SiteEmissions = CUE * ICTEnergy \quad (4)$$
$$= PUE * ESC * ICTEnergy. \quad (5)$$

B. Algorithm description

This subsection describes the functionalities of the default round-robin cluster selection algorithm, as well as the two developed algorithms for optimizations: Fastest possible (FP) that tries to minimize the wait time, and Energy and CO_2-aware (ECA) that tries to minimize the energy or CO_2 emissions.

1) Round-robin (RR): RR algorithm is generally used in USO for selecting the cluster. It balances the number of jobs between different clusters by always choosing the next cluster compared to the previous selection. After the last cluster, the selection is started again from the first cluster.

2) Fastest possible (FP): FP cluster selection algorithm tries to select the cluster that could possibly execute the job with minimal wait time. For this, the algorithm first checks if there are enough idle nodes/cores in some cluster for executing the job. If yes and the cluster's queue is also empty, the job is submitted to that cluster. If not, an estimated wait time for the job in each cluster is calculated by using the current status of each cluster: number of nodes and cores, status of running jobs, number of jobs in the queue, and walltimes of each queued job. The cluster with the shortest estimated wait time is then selected.

The algorithm relies on the dynamic cluster properties (status of nodes and queues), which can be obtained by a single site monitoring system. Otherwise, this dynamic information is not available for the USO, so the normal cluster selection algorithms can exploit only static cluster information for the decision making.

It should be noted that the wait time can only be estimated. The walltimes of the jobs are given by the users and, in general, they are inaccurate [28], [29]. Also, the used scheduling algorithm affects in which order the jobs are executed (especially backfilling). Thus, it is possible to calculate only the maximum wait times for the jobs, not the exact wait times.

3) Energy and CO_2-aware (ECA): This algorithm tries to find the cluster with the smallest amount of estimated energy consumption or CO_2 emissions; the optimization goal can be chosen by the user. The CO_2 emissions of the job can be calculated in the same way as for the whole site in Equation (4):

$$JobEmissions = CUE * ICTEnergyOfTheJob. \quad (6)$$

The simplest way is to select the cluster with the smallest CUE value. This works if the clusters have significant differences in their CUE values (CUE = ESC * PUE). If there are only small differences in the CUE values, then additional estimations should be done, since the job may consume different amount of ICT energy in different clusters due to the different computing node properties (CPU, RAM, etc.), and this difference may become a greater factor than CUE for the CO_2 emissions. The ICT energy of the job can be estimated by using the job requirements (number of nodes/cores, walltime) and cluster's computing node properties (CPU, RAM, etc.) as inputs for power consumption models such as those described in [2] and [30]. Moreover, the job execution time may differ greatly between clusters because of

varying hardware parameters. Thus, it is very important to try to estimate the execution time of the job on different clusters. Special application benchmarks for estimating the clusters' job execution times were chosen for this purpose and they were used in the testbed experiments. These benchmark applications are run once on each cluster beforehand, and the execution times are measured. Based on the measurements, the data centre operators rank each cluster and benchmark application combination. The higher the rank, the faster the execution time is.

Thus, the ICT energy of the job is given by

$$ICTEnergy = ICTPower * ExecutionTime, \quad (7)$$

where $ICTPower$ is the power consumption of the job and is calculated by using the power consumption models, and $ExecutionTime$ is calculated by using the walltime estimate and the rank of the cluster.

However, selecting always the cluster with the least amount of estimated CO_2 emissions would cause huge load and queue on the cluster with the least CO_2 emissions. This would mean large delay for the users. Thus, some form of load balancing is needed for this algorithm. In the conducted simulations (described in the next sections), we used a queue size limit: If the queue exceeded its size limit, the job was submitted to the cluster with the second least CO_2 emissions, and so on. In the case of green SLA, the users set a certain deadline for the completion of their job. This limit can be used for load balancing: The estimated completion time for the job can be calculated as a sum of the estimated wait time and walltime of the job. If this is in the limits, the cluster can be chosen. If not, the same calculations should be made to the cluster with the second least CO_2 emissions, and so on. If the user sets too strict time limit for the job that none of the clusters can fulfill, the job should be either denied or the cluster should be chosen by the Fastest possible algorithm. In the testbed experiments (described later) we used this green SLA method for load balancing, and FP algorithm in case of too strict time limits. The ECA algorithm can be used for selecting the cluster with minimal energy consumption, too, by replacing CUE by PUE.

The ECA algorithm takes into account the dynamic properties of the cluster and compute nodes. This information is stored in a meta-model, which is updated accordingly if any of the parameters, such as PUE, CUE or compute node hardware parameters are changed.

IV. Simulation studies

The simulation model for the HPC grid has been developed with the OMNeT++ discrete event network simulator [31] and the INET Framework [32]. The design of the model is similar as in [2], except that the model is extended from a single site scenario to a federated site scenario.

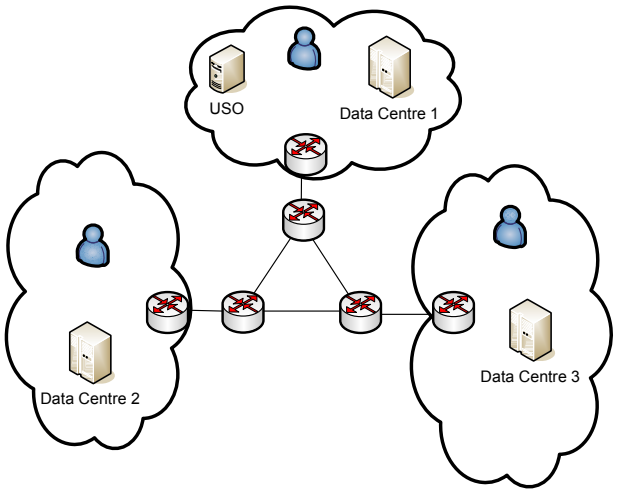


Figure 2. Network topology

Figure 2 illustrates the network topology used in the simulations. It consists of three backbone routers, three gateway routers, three data centre modules, three clients and a USO module. In this scenario, the clients send HPC job requests to the USO, which is responsible for choosing an appropriate data centre, i.e., an HPC cluster, for executing the job. The USO has been adapted for the simulation so that it is capable of using the developed optimization algorithms and making decisions based on the dynamic properties of the clusters. Normally, only static information of the clusters is available for the USO.

For the decision making, the USO can query the status and properties of each cluster from the corresponding RMS. Once the cluster is chosen, the USO forwards the job request to the RMS of the chosen cluster. The RMS uses the policies and scheduling algorithms of the cluster to choose suitable servers for job execution. When the job execution finishes, the RMS informs the USO, which again forwards the information to the client that submitted the job for execution.

The data centre module can be seen in Figure 3, which is similar as in the single site scenario used in our previous work [2]. It contains a RMS, a fixed number of servers and a router between them. The RMS handles all incoming job requests arriving to the data centre and allocates the jobs to the servers for execution according to the selected policies and algorithms. Thus, the RMS also functions as a scheduler in the simulation. The RMS supports 6 different scheduling algorithms: standard FIFO, Backfill First Fit and Backfill Best Fit algorithms and their energy-aware counterparts developed previously (see [2] for more information on these).

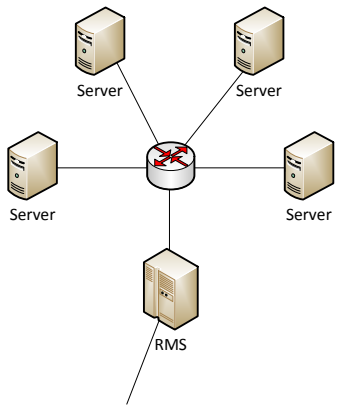


Figure 3. Data centre module

The RMS module includes parameters for the PUE and the CUE. By using these two values, the USO is able to select a cluster that is the most energy-efficient or produces the least amount of CO_2 emissions.

A. Simulation scenario

In this subsection, we describe the simulation scenario and parameters. For evaluation we consider a scenario that includes three data centres and 75 clients that are sending job requests to the USO. The simulation is stopped once 1500 jobs have been completed. During the simulation we measure the energy consumed by each data centre and present the obtained results in the next section. General simulation parameters are presented in Table II. Uniform(a,b) means randomly selected value according to a uniform distribution between a and b. The scheduling and cluster selection algorithms are shortened as follows:

- FP = Fastest possible USO cluster selection algorithm
- ECA = (Energy and) CO_2-aware USO cluster selection algorithm set to minimize the CO_2 emissions
- RR = Round-robin USO cluster selection algorithm
- FIFO = First In, First Out job scheduling algorithm
- BFF = Backfilling first fit job scheduling algorithm
- BBF = Backfilling best fit job scheduling algorithm
- E-FIFO, E-BFF, E-BBF = energy-aware counterparts for the job scheduling algorithms (idle nodes are powered off whenever possible)

In Table III, we can see the parameters for the three clusters in the considered federated HPC data centre. The clusters have different characteristics, such as, the number of servers, PUE, and ESC. The energy sources (O = Oil, C = Coal, H = Hydro, N = Nuclear) for the clusters were selected so that both extreme ends in terms of ESC were represented in the simulations, while the third one represents something in the middle of them. Also, servers have different operating systems (OS) and processor architectures. In the simulations, the ECA algorithm optimization goal was set to minimize the CO_2 emissions.

Table II
Simulation parameters

Parameter	Value
Simulation runs	10
Number of jobs	1500
Number of data centres	3
Number of clients	75
Number of gateway routers	3
Number of backbone routers	3
USO cluster selection algorithm	RR, FP, ECA
RMS scheduling algorithm	FIFO, BFF, BBF, E-FIFO, E-BFF, E-BBF
Server memory	4 * 2 GB = 8 GB
Server cores per CPU	2
Server CPUs	2
Server CPU idle power	15 W
Server core voltage	1.2 V
Client job cores	1, 2, 4
Client job load	Uniform(30,99)
Client job nodes	Uniform(1,20)
Client job memory	Uniform(100MB, 2GB)
Client job run time	Uniform(600s, 86400s)

Table III
Data centre parameters

Parameter	Cluster 1	Cluster 2	Cluster 3
Servers	30	40	50
Energy source	C 50% H 20% N 30%	C 80% O 20%	O 20% H 40% N 40%
PUE	1.5	1.8	1.3
ESC	0.45983	0.85	0.12844
CUE	0.689745	1.53	0.166792
OS	Linux	Windows	Linux
CPU arch.	AMD	Intel	Intel

B. Simulation Results

Figure 4 presents the total ICT energy consumption of the three clusters for different USO cluster selection and job scheduling algorithms. As can be seen, RR with normal job scheduling algorithms consumes the most amount of energy. RR with normal job scheduling algorithms represents a generally used, un-optimized algorithm combination in federated HPC data centres. Thus, it serves as a comparison point when calculating the energy savings and CO_2 emission reductions.

Figure 5 presents the energy savings achieved by using Fastest possible and CO_2-aware USO cluster selection algorithms instead of the default RR algorithm, and by using energy-aware job schedulers on each cluster. The energy-aware job schedulers are compared to their normal counterparts; for example, the first bar (E-FIFO FP) means the savings compared to FIFO RR. The last three bars present the savings when using RR but with energy-aware job scheduling. It can be seen that by using energy-

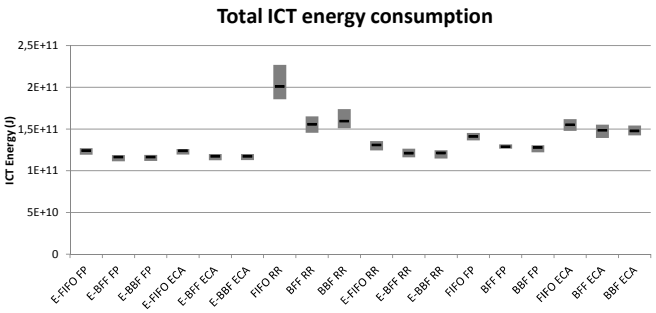


Figure 4. Total ICT energy consumption. Black lines represent the average value and the floating bars show the range of values from minimum to maximum

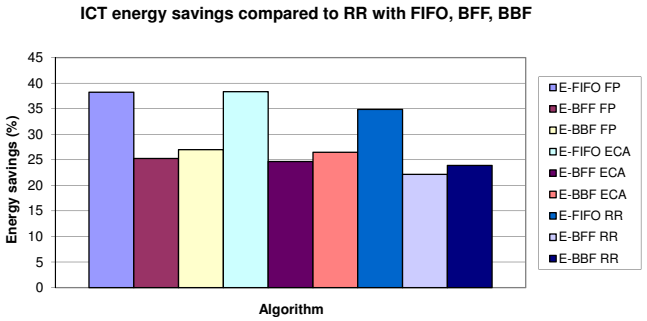


Figure 5. ICT energy savings compared to un-optimized, generally used RR with FIFO, BFF, and BBF

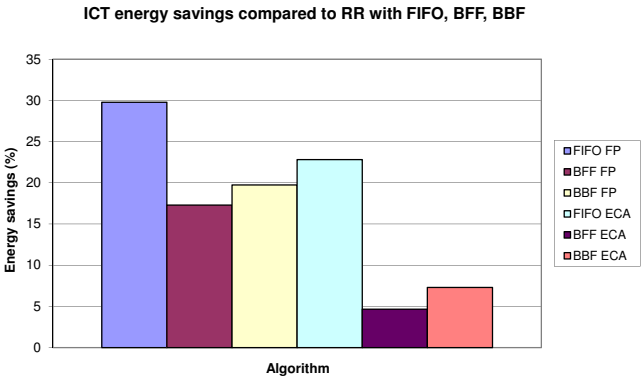


Figure 6. ICT energy savings compared to RR with normal job scheduling

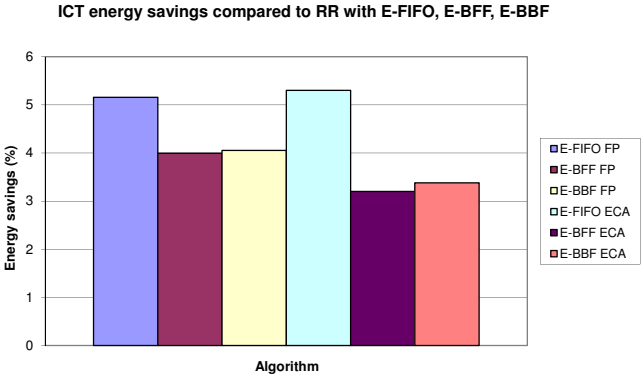


Figure 7. ICT energy savings compared to RR with energy-aware job scheduling

aware job scheduling, 22% to 35% energy savings can be achieved. Together with FP and ECA cluster selection, the savings are about 25% to 38%. When comparing the energy-aware algorithms to their normal counterparts, the savings with E-FIFO are the largest. This is because backfilling exploits the idle nodes more efficiently by running shorter jobs while with standard FIFO the nodes that cannot be used for job execution are left in an idle state and can thus be shut down by the energy-aware scheduler.

However, if we only change the cluster selection algorithm, and keep the normal job scheduling algorithms, we can see from the Figure 6 that with FP we can save 17% to 30 %. Since the cluster selection is performed before job scheduling, we can say that about 8 % of the total savings are due to the energy-aware job scheduling, while the rest is due to the FP cluster selection. When comparing to RR with energy-aware job scheduling (as depicted in Figure 7), we can see that FP and ECA cluster selection algorithms can save additionally about 3% to 5%. For the explanation, we have to take a look at the jobs' average wait and turnaround times and the simulation duration.

Figure 8 presents the average wait times of the jobs (i.e., the average waiting times of the jobs in the queue) in case of different USO cluster selection and job scheduling algorithms. As can be seen, the average wait time is clearly shorter with the FP USO algorithm. The ECA USO algorithm with backfilling has about the same average wait time as RR, even though RR with FIFO clearly has the longest waiting time. Also, there are basically no differences between RR with energy-aware and normal job scheduling. This is true also in general, as reported in [2]: energy-aware job scheduling does not cause significant increase in wait time.

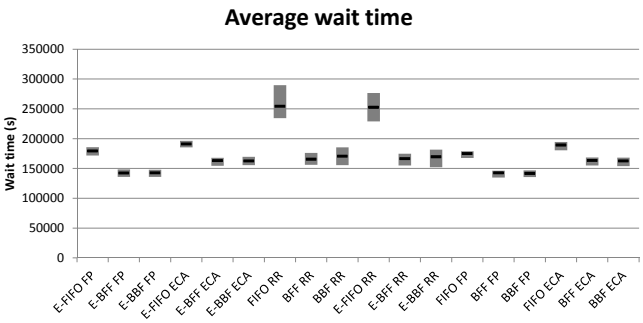


Figure 8. Average job wait times

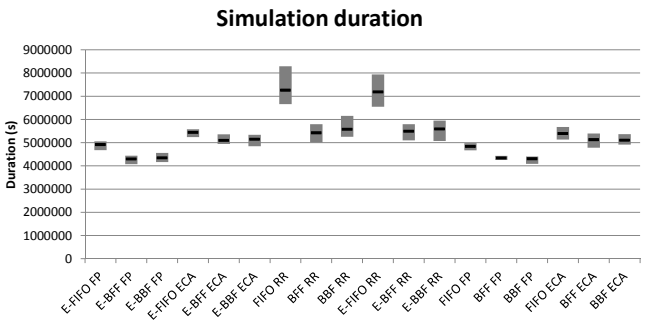


Figure 9. Average simulation duration

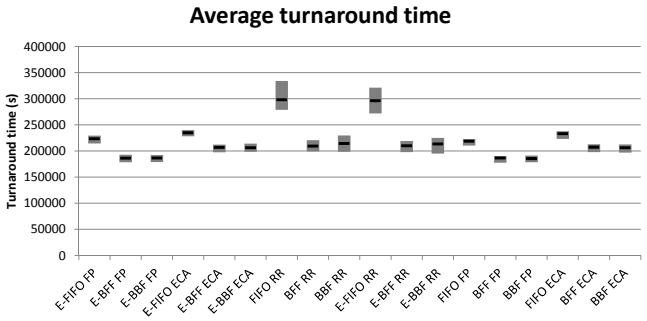


Figure 10. Average job turnaround time

Figure 9 depicts the simulation duration, i.e., how long time it took to execute all the 1500 submitted jobs. The graph shows the same as Figure 8: because the wait times are longer with RR USO cluster selection, also the simulation duration is longer.

Figure 10 presents the average job turnaround times in case of different scheduling algorithms. The story is the same as in previous figures: RR is slower due to the longer wait time.

Based on the results above, we can conclude that RR cluster selection with normal job scheduling algorithms can be very inefficient in terms of energy. This is because RR only balances the number of jobs among the clusters. It does not take into account the differences in the clusters (e.g., number of nodes/cores) or the differences in the submitted job characteristics (e.g., number of nodes/cores, walltime estimate). This can lead to a situation where one cluster is over utilized with many jobs waiting in the queue, while the other clusters can be under utilized at the same time, with nodes running idle. The energy-aware job schedulers (E-FIFO, E-BFF, E-BBF) power off the idle nodes whenever possible, and this is why a substantial amount of energy can be saved. On the other hand, the FP cluster selection algorithm inherently takes into account the differences in the clusters and submitted jobs: it always selects the cluster with the estimated minimal wait time, and thus balances the utilization between the clusters. Then fewer nodes are running idle

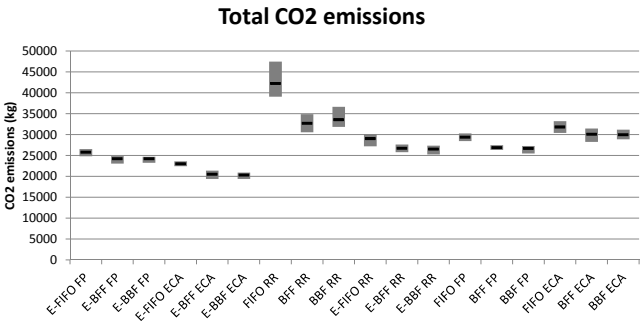


Figure 11. Total CO_2 emissions

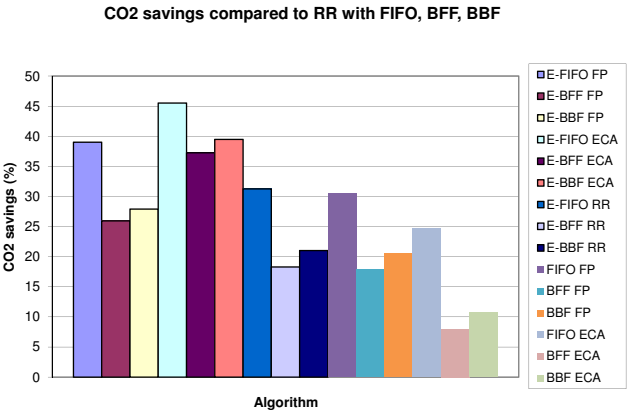


Figure 12. CO_2 savings compared to RR with FIFO, BFF, and BBF

and energy is saved. Also ECA saves some energy even though its goal was set to minimize the CO_2 emissions.

Figure 11 presents the total CO_2 emissions of the federated HPC data centre. As can be seen, RR with normal job scheduling causes the largest CO_2 emissions. Using energy-aware job scheduling reduces the emissions due to the reduced energy consumption. Using FP cluster selection reduces the energy consumption still a bit more due to the better load balancing among clusters, and thus the CO_2 emissions are also smaller. ECA cluster selection algorithm favours the cluster with the best CUE value, i.e., least amount of CO_2 emissions, and hence achieves the greatest savings in CO_2 emissions, about 37% to 45% compared to RR with normal job scheduling. Note that this requires also using the energy-aware job scheduler; without energy-aware job scheduling the emission reductions are smaller since ECA prefers Cluster 1 over Cluster 2 due to the smaller CUE, which in turn results in worse utilization in the bigger Cluster 2. Energy-aware job scheduling turns off the idle nodes on Cluster 2 and hence cuts the emissions. The CO_2 savings are depicted in Figure 12 as percentages.

V. Testbed experiments

The energy-aware job scheduling and cluster selection algorithms were also implemented as part of the software plug-in developed within the project FIT4Green [33]. The developed plug-in [34] is a set of software components that add energy management capabilities to a data centre by interfacing with the existing management and automation tools of the data centre. Its goal is to dynamically optimize the deployment of the applications and services hosted and running across the ICT resources of a single or federated site data centre, in order to minimize the energy consumption or CO_2 emissions, that is, trying to consolidate load so as some hardware resources can be turned off or set to low-power state. One core software component of the plug-in is the *Optimizer* that contains the cluster selection and job scheduling algorithms and uses them together with the information on data centre resources for suggesting a list of actions that can potentially lead to reduced energy consumption.

In particular, the plug-in communicates with the RMS and UNICORE at the HPC data centre environment by acquiring information about the data centre status: the jobs in the queue, status of the nodes and the jobs that are currently running. All this information is stored and kept up-to-date in an XML-based meta-model. Based on this information, the plug-in can send actions to the RMS and UNICORE, such as, start job or shutdown node in the single site scenario. In the federated scenario, the plug-in decides to which cluster the job will be submitted to.

The plug-in can be used also in other types of data centres, i.e., in traditional service or enterprise portals, or cloud computing data centres. However, these are out of scope of this work. An interested reader can find more detailed information about the plug-in in [34]. The plug-in's source code is available at [35] as open source licensed under the Apache License, Version 2.0.

A. Testbed scenario

This subsection describes the testbed scenario: the clusters used in the tests, their configuration and characteristics, how the tests were conducted and what kind of test workload was used.

1) Environment and configuration: The testbed scenario consists of three HPC clusters: Juggle, Jufit and Dune. Juggle and Jufit are located at the Jülich Supercomputing Centre in Jülich, Germany, while Dune is set up at the VTT Technical Research Centre in Oulu, Finland. By having three testbed clusters located at different sites and countries, it is possible to analyse the impact of different CUE and PUE parameters of real distributed systems. In Juggle and Jufit it is possible to set the compute nodes to a low-power standby mode, while in Dune it is only possible to shut down nodes.

Table IV
Operating numbers of the Juggle cluster

Parameter	Value
Processor type	Dual AMD Opteron F2216 2.4GHz
Number of nodes	1 head node, 12 compute nodes
Cores per node	4
Overall number of cores	48
Main memory	8 GB per node
Network	InfiniPath(QLOGIC), Gigabit Ethernet
2 file servers	disk capacity: 6 TB
Power supply efficiency	83%
Operating system	SLES 10, Scientific Linux 5.2
RMS	Torque (PBS Scheduler)
Node power consumption	
- standby	117W
- idle	162.5W
- maximum	230W

Juggle involves altogether 12 compute nodes (see Table IV). This allows executing jobs requiring many resources in parallel. The Juggle is a relatively old system compared to the other clusters in the testbed. Intelligent Platform Management Interface (IPMI) services and monitoring tools if not already provided by the operating system have been installed on each node of the system to enable the monitoring of dynamic system parameters, such as core voltage and frequency, memory load, fan RPM, and disk read/write rates.

Jufit is a more modern system providing a more modern generation of processors. It consists of two compute nodes with 12 cores each (see Table V). Compared to Juggle, Jufit is in general more energy efficient in terms of CPU power consumption and power supply efficiency. The Jufit cluster has been used in the measurements for the federated scenario.

The test environment at VTT consists of a Linux cluster framework called Dune that includes four compute nodes and a head node. All nodes are rackable Dell PowerEdge R510 servers with equal characteristics as can be seen in Table VI. There is no separate file server available in the cluster, but instead all the nodes hold adequate hard disk drives, with 1 TB of disk space.

The Torque RMS is installed on all clusters for managing nodes and the scheduling and monitoring of jobs. Furthermore, Target System Interface (TSI) modules of the UNICORE middleware are installed on the head nodes of the clusters to allow submitting jobs by UNICORE, which is needed in the case for the federated scenario.

All clusters are connected to Power Distribution Units (PDUs), which measure the power consumption of each single head and compute node. The results are requested conveniently by clients through the Simple Network Management Protocol (SNMP). The measurements were

Table V
Operating numbers of the Jufit cluster

Parameter	Value
Processor type	Quad-core Intel Xeon X5660 (Westmere), 2.6 GHz
Number of nodes	1 head node, 2 compute nodes
Cores per node	12
Overall number of cores	24
Main memory	24 GB per node
Network	InfiniPath(QLOGIC), Gigabit Ethernet
2 file servers	disk capacity: 6 TB
Power supply efficiency	91%
Operating system	OpenSuSE 11.3
RMS	Torque (Maui Scheduler)
Node power consumption	
- standby	142W
- idle	175W
- maximum	232W

Table VI
Operating numbers of the Dune cluster

Parameter	Value
Processor type	2 x Quad-core Intel Xeon E5606 (Westmere), 2.13 GHz, 64-bit
Number of nodes	1 head node, 4 compute nodes
Cores per node	8
Overall number of cores	32
Main memory	4 x 16 GB (aggregate 64 GB)
Network	Gigabit Ethernet
Power supply efficiency	90%
Operating system	64-bit Rocks cluster distribution (based on CentOS 5.6 Linux)
RMS	Torque (Maui Scheduler)
Disk space	1 TB SATA HDD on each node
Node power consumption	
- off	0-2W
- idle	85-90W
- maximum	165-175W

updated every 3 seconds during the tests, so that the values could be provided reliably.

The motherboards of the Juggle and Jufit compute nodes support the ACPI [36] state S1, which is also known as 'standby' state. As soon as the software plug-in is generating a standby action, the appropriate compute node will be set to standby mode. While on Juggle 'wake-on-lan' is used to bring the machine back from standby to normal state, the same result on Jufit is achieved by an IPMI wake up command. The VTT testbed cluster Dune is using instead the ACPI state S5 or better known as 'soft-off', which requires a full reboot of the system.

Additionally, a DVFS feature has been enabled on Juggle to analyse the impact of using DVFS instead of ACPI energy saving mechanisms.

2) Testing methodology: The goal of the test measurements was to compare the energy consumption of the plug-in adapted supercomputing environments with systems using default state-of-the-art solutions. The tests tried to analyse the energy saving capabilities of the plug-in software in two different areas:

- Savings on a single cluster by using the energy-aware job scheduler of the plug-in
- Savings in a federated cluster scenario by using the cluster selection algorithms of the plug-in

For each of these investigation areas specific test approaches were carried out regarding the used benchmark workloads, type of job submission, and energy metering. All measurements depended on the available hard- and software. The workloads stressing the testing environment consisted of jobs that made use of the installed HPC applications. The jobs were either submitted directly from the test user's home directory on the head node of the cluster or alternatively from the UNICORE client, so that the user did not need to be logged in to the cluster. Both submission types are quite common in supercomputing scenarios.

The single site scenario tests were performed on the Juggle system, which provides a suitable number of nodes for testing the energy saving potential by using the plug-in's energy-aware job scheduler. Firstly, this scheduling mechanism schedules jobs in the queue of the cluster to the particular nodes and cores of the cluster, and, secondly, it sets nodes to standby if they are completely idle. Alternatively, DVFS instead of ACPI standby can be used to save energy. When having equal workloads and comparing the energy-aware scheduler with a default state-of-the-art scheduler, the best possible energy consumption depends on how efficiently the jobs can be scheduled without loss of time, so that as many nodes as possible can be set to standby. In the supercomputing testbed at FZJ the energy measurements were analysed by comparing the default PBS scheduler with the plug-in's energy-aware scheduler. While the PBS scheduler is based on an enhanced FIFO (first in, first out) algorithm, the plug-in one used the backfill first fit approach. The particular measurements were performed with workloads generating different system utilization profiles on the Juggle cluster (0%, 40%, 60%, and 80%) to simulate potential loads of real supercomputing machines.

The total energy consumption generated by a single test workload has been calculated in Joule as a product of the measured average power of all cluster nodes and the elapsed time that was needed to run all jobs of the workload. The elapsed time involves the total time elapsed from the submission of the test user's first job until the output files of the last executed job have been stored where requested. So, this period includes the time for transferring input files, the wait time in the RMS queue, the actual execution time, and the time to stage the output files to the requested locations. Each measurement was stopped immediately after all jobs

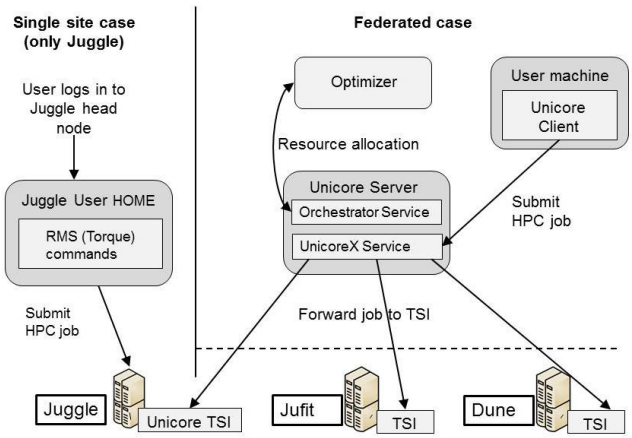


Figure 13. Difference of job submission in single and federated benchmark tests

of the test workload were finished. This approach was mandatory to measure the time that different scheduling strategies need to process the workload. So, the energy of one measurement was calculated as follows:

$$E_{SingleSite} = T_{Workload} * P_{Cluster}, \quad (8)$$

where $E_{SingleSite}$ is the total energy of the site, $T_{Workload}$ is the elapsed time to process the workload, and $P_{Cluster}$ is the average power consumption in the cluster during the workload processing.

Single site scenario tests were also performed on the Dune cluster following the same testing methodology. The only differences were that Dune supports only soft-off instead of standby and DVFS, and the comparison point was Maui scheduler instead of PBS scheduler.

When performing tests in the single site scenario, the jobs of the workload were submitted locally from the cluster's head node by using provided shell commands of the RMS on the dedicated cluster. This approach is a common practice in supercomputing environments when the dedicated target system of the job is already known. In the federated scenario another job submission approach was utilised. Since it is not known in advance on which cluster the job should be executed, the test user submits the jobs at first to a UNICORE server, which acts as a centralized entry point for all incoming jobs. This UNICORE instance is connected to the installed UNICORE TSIs on the testbed clusters, so that the server is able to submit incoming jobs to the RMS of an appropriate machine. Before that, the USO, service of the UNICORE server, initiates a resource allocation request to ask the plug-in for a suitable target machine for the job. Figure 13 highlights the difference in terms of job submission between the single and federated scenario.

In general, the workloads have been compared in each test case by using the plug-in's scheduling strategies as well as default mechanisms for getting results of how efficiently the plug-in is able to save energy. In the case when the plug-in was not used, the existing state-of-the-art mechanisms have been used to process the jobs on the test clusters.

The approach of the energy consumption measurement in the federated scenario is to consider only the elapsed time, which is needed on each testbed cluster to run the assigned jobs of the benchmark workload. This incorporates that each involved cluster can produce in one benchmark measurement a different elapsed time and different average power consumption. So, the total energy consumption $E_{Federated}$ in one measurement is the product of the elapsed time and the average power of each cluster i:

$$E_{Federated} = \sum_{i=1}^{N} T_{Cluster_i} * P_{Cluster_i}. \quad (9)$$

3) Test workload: The benchmark measurements on the testbed were aimed at stressing the testbed as close as possible to clusters in real supercomputing environments. For that purpose different typical HPC applications were installed on the test clusters, namely LINPACK [37], a collection of FORTRAN subroutines to solve linear systems, and PEPC (Pretty Efficient Parallel Coulomb Solver) [38], which is used to run astrophysical N-body simulations.

For the single site scenario the test workloads were created by a configurable Perl script, which can parameterise the jobs in terms of the used HPC application, the level of computation intensity, the number of used nodes and cores, as well as the planned walltime (also known as wall clock time), which is the time elapsed until a job should have been finished. In this way workloads were created stressing the system with different system utilization in order to analyse the energy savings under those different loads. Also real world clusters working in production show often varying system utilizations between entirely idle and working to capacity. The utilization factor is defined here as the percentage of time when cluster resources are stressed with jobs relative to the total elapsed time of the workload. For instance, a system load of 90% means that only on an average of 10% of the elapsed time cluster nodes are able to be set to the energy saving standby mode because they are otherwise busy with running jobs.

In case of the federated scenario the workloads of the single site scenario were adapted as a template to create workloads using the same HPC applications within the graphical UNICORE Rich Client (URC) [39]. This workload is embedded in a workflow from where the single jobs are submitted in parallel to the UNICORE server, which in turn initiates resource allocation requests to the plug-in's cluster selection algorithms, and forwards

Table VII
DUNE - NUMERICAL RESULTS OF SINGLE SITE MEASUREMENTS

System utilization [%]	0%	50%	80%	90%
Energy with plug-in [kJ]	11.8	133.2	433.4	480.5
Elapsed Time [s]	517.8	543.2	1276.9	1276.7
Average Power [W]	22.8	250.8	339.6	376.6
Energy without plug-in [kJ]	176.9	254.5	499.1	515.9
Elapsed Time [s]	515.7	648.0	1258.3	1268.0
Average Power [W]	343.4	393.0	396.8	406.9
Energy saving [%]	93.3	47.7	13.2	6.9

subsequently the jobs to the chosen cluster.

Roughly 90% of the workload jobs can run on both clusters, while the requirements of always 10% of the jobs can only be met on one of the clusters. This distribution creates a base load on the clusters to map better real world environments where jobs are usually not able to run on every available target machine.

B. Testbed results

This subsection presents the results of the testbed experiments for both single and federated site scenarios.

1) Single site scenario: In our previous work [2], we showed the potential energy savings in a testbed that considered the ACPI standby mechanism on the compute nodes. In this work, we have performed additional single site measurements on the new VTT testbed cluster Dune, which is using ACPI S5 soft-off power shut down to save energy on unused nodes, while on Juggle and Jufit only ACPI standby is available. ACPI standby is faster than soft-off in switching to the power-on state but, on the other hand, does not have the same power saving potential. The results can be found in Table VII. Each workload measurement was repeated with several iterations.

From the results it is possible to note that the energy savings are highly dependent on the test workload and the utilization of the whole system. The single site optimization algorithm attempts to keep the system active with the lowest possible amount of resources and still providing suitable turnaround times for the jobs. With lower utilization values the energy savings are higher, but if the system is very busy there are not many opportunities to shut down idle resources.

The single site scenario tests at FZJ were performed on the Juggle system, which provides a suitable number of nodes for testing the energy saving potential by setting nodes to standby status. However, apart from the different ACPI mechanisms it was worth to evaluate the energy saving potential of using the DVFS feature, which works on the principle of setting unused nodes to the powersave governor, i.e., the lowest CPU frequency and core voltage is set on the node. The performance governor is set again by the plug-in once the nodes are requested again by jobs. That setting implies that the maximum available frequency is set on the CPUs of the nodes. On-demand governors were consequently not used in the configuration, since it is in general not desirable to use frequency scaling when running CPU intensive jobs. Administrators experienced performance drawbacks with on-demand governors, which is a critical issue in HPC environments. The software plug-in on Juggle has thus been enabled to make use of different energy saving methods:

- ACPI power saving statuses (e.g., standby or soft-off)
- DVFS (switching between powersave and performance governor)
- ACPI standby + DVFS (observed slightly higher savings on the testbed cluster when using both mechanisms in parallel).

In general, the energy measurements have been analysed by comparing the PBS default scheduler with the the plug-in's energy aware job scheduler. These measurements were performed with workloads generating a different total load on the cluster (0%, 40%, 60%, and 80%). Each workload measurement was repeated with several iterations. The results in Table VIII show the average values of those tests.

The energy consumption of a single workload has been calculated in Joule as a product of the measured average power of all cluster nodes and the elapsed time, which was needed by the appropriate workload. The elapsed times of each workload depend on the composition of the workload to achieve certain system utilization, so there is no correlation between the elapsed time values of different system loads. In contrast, the average power consumption increases with more intensive workloads, since less idle nodes can be set to an energy saving status.

When comparing the values of default scheduler measurements with energy-aware scheduler tests it is apparent that the elapsed time is most time slightly higher with the energy-aware strategies, which is caused by the overhead of the optimization process. In particular, cluster information such as node and job statuses must be read and analysed at the remote plug-in server, and generated actions must be sent back to the RMS of the cluster. However, this process has been optimized in terms of performance during the implementation phases, so that the time impact has been minimized. Overall, the expected overhead in the single site scenario is round about 2-3% compared to default mechanisms.

Taken into account similar elapsed time results for

Table VIII
JUGGLE - NUMERICAL RESULTS OF SINGLE SITE MEASUREMENTS

System utilization [%]	0%	40%	60%	80%
Energy with standby [kJ]	1300	2652	3018	4410
Elapsed Time [s]	1000	1634	1580	2143
Average Power [W]	1300	1623	1910	2058
Energy with DVFS enabled [kJ]	1400	2701	3033	4433
Elapsed Time [s]	1000	1556	1558	2126
Average Power [W]	1400	1736	1947	2085
Energy with standby+DVFS [kJ]	1280	2616	3003	4389
Elapsed Time [s]	1000	1620	1579	2144
Average Power [W]	1280	1615	1902	2047
Energy with default scheduler [kJ]	1960	3304	3345	4625
Elapsed Time [s]	1000	1554	1552	2088
Average Power [W]	1960	2126	2155	2215
Energy saving with standby [%]	33.7	19.7	9.8	4.6
Energy saving with DVFS [%]	28.6	18.2	9.3	4.2
Energy saving with standby + DVFS [%]	34.7	20.8	10.2	5.1

the single site measurements, the main factor for saving energy is clearly the measured average power of the tested cluster, which is proportional in the measurements to the decreasing system load. While the default RMS scheduler cannot make advantage of idle compute nodes, the plug-in sets them in an energy saving standby mode, which reduces noticeably the average power of the system.

When using only DVFS we could observe energy saving between 28.6% and 4.2% compared to the default scheduler depending on the generated system utilization. When enabling DVFS and ACPI standby together we could even achieve between 34.7% and 5.1%, which confirmed the assumption that DVFS + standby generates a slightly lower consumption than ACPI standby alone (between 33.7% and 4.6%). However, this behaviour could only be detected on the used testbed hardware. It cannot be stated as a general rule. Nevertheless, administrators can check their systems if it makes sense to use both mechanisms in parallel. On Juggle an additional gain of about 1% saving is possible when using both features. Using DVFS alone could be a benefit when high job fluctuations can be expected on a system. Especially, when the hardware is supporting only a full shut down ACPI status, which would need 2-3 minutes for powering the system on again, and a high job submission rate is supposed, it could be more efficient to use the supported DFVS feature. That mechanism needs only one second to switch between the governors.

2) Federated scenario: In the supercomputing federated scenario we wanted to measure the emission and energy saving capabilities of the plug-in's cluster selection strategies. Jobs should be assigned to suitable cluster resources in the most energy efficient way. The plug-in's Optimizer implements two different strategies to achieve that resource allocation, namely the ECA and FP cluster selection algorithms.

Section III gives already a detailed description of both strategies. In short, the FP algorithm calculates the wait time of a job on all potential suitable clusters and submits the job to the system that provides the most minimal estimated queue time. In contrast, the ECA algorithm estimates at first the energy or respectively the emission (depending on the chosen objective), which would be produced by a job on a particular cluster. The energy/emission is calculated by considering the PUEs/CUEs of the clusters as well as estimating the ICT energy consumption of particular. In the supercomputing testbed at FZJ, PUE and CUE indexes are equal for both testbed clusters, since both machines are located in the same data centre environment. However, the Dune cluster of VTT is located at Oulu in Finland and provides different PUE and CUE specific values.

Additionally, the ECA algorithm checks if the user defined 'latest job finishing time' can be satisfied, i.e., we used user defined allowed delay value for load balancing between clusters. This means that the plug-in verifies if the job can be executed and finished on a cluster within a user defined time limit. If not, the job cannot be scheduled in the best energy efficient way and the next cluster is chosen, where the estimated wait time is smaller than the user defined value. By this mechanism, the user can set a threshold value from where jobs must not be scheduled energy efficiently anymore.

The crucial factor in terms of energy efficiency is the power consumption of the testbed clusters over a dedicated time. Furthermore, it was worth to analyse the impact of an application benchmark parameter, which takes into account the different performance of an application produced on a particular cluster.

Scheduling jobs in a federated environment of different supercomputers is not yet very common in HPC. Resource brokering in heterogeneous environments is still an on-going research area in HPC. Meta-schedulers are usually not yet deployed in real production environments but rather in smaller research and test environments. One of the main barriers is that HPC job requirements are often strongly system-related so that the jobs can only run in a dedicated hard- and software environment. There is also a lack of dynamic resource information

on the broker level about node and queue statuses of the bounded supercomputers. Usually, users have a clear picture where to submit their jobs. Often these jobs can only run on particular nodes, since only there the required application software is installed.

Accordingly, there are no matured solutions of meta-schedulers available that could be used as reference software. So, in the testbed experiments we compared our solution with a simple round-robin scheduling mechanism provided by the UNICORE software. This solution does not consider any dynamic system information about jobs in queues or the statuses of nodes. A more intelligent solution for scheduling jobs in federated environments is the developed plug-in's FP algorithm, which does not take into account energy or emission aware parameters but only the fastest execution time of jobs on the clusters. So the focus of the federated scenario evaluation was not only to compare ECA and FP with default RR algorithm, but to check the impact of the provided scheduling parameters on the results. In the federated tests the energy-aware job scheduling used in the single sites assessment was deactivated to set the focus on the energy saving potential of the federated algorithms.

For analysing the impact of different PUEs on the sites we changed the PUE value of the FZJ site in each trial and used a fixed value on the VTT site where Dune is located, so that we could analyse the system utilization of the clusters while changing the PUE difference of them. The PUE value on Dune, which was calculated by VTT administrators, was set to 1.2. For the FZJ site we actually calculated a value of 1.4. In the evaluation we iterated the FZJ parameter from 1.0 to 1.8.

Figure 14 shows clearly the impact on the utilization of the particular testbed clusters when iterating over the FZJ PUE value. The graph of the Juggle is relatively constant, since the Optimizer schedules only some default jobs to the system, which can run only there (requesting 8 nodes per job). Apart from these jobs no others are submitted to that machine, since the Optimizer considers its high basic power consumption compared to the other two clusters. In contrast, Jufit is getting at first the most jobs of the workload when having a lower PUE than Dune which was set to 1.2. The consequence is at the beginning higher system utilization on Jufit compared to Dune.

When Jufit and Dune have the same PUE (1.2) Dune is already preferred clearly by the Optimizer since it has the most efficient power consumption per compute node in the testbed. Higher PUEs for FZJ-Jufit result in even lower utilizations for Jufit and higher ones for Dune. With more than 1.4 PUE on Jufit saturation on the utilization of Dune can be detected. Dune is utilized almost with 100% and also Jufit levels out at about 35%.

The CUE evaluation was performed in a similar way

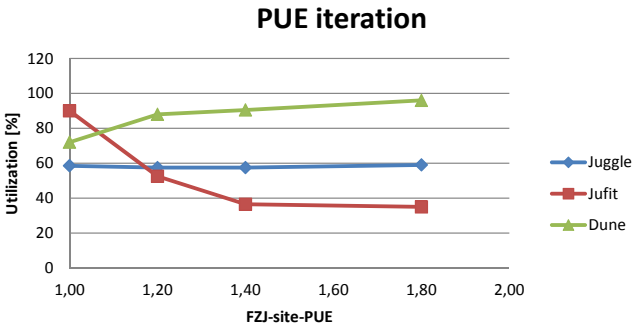


Figure 14. Impact on cluster utilization when iterating over PUE

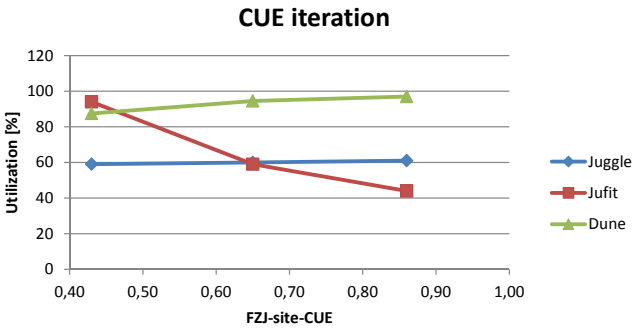


Figure 15. Impact on cluster utilization when iterating over CUE

as the PUE iteration. After changing the Optimizer objective to CO_2 emissions, the CUE parameter of the available two sites was used for calculating the job scheduling in the federated scenario tests. For Dune a CUE value of 0.43 was calculated based on local energy and emission characteristics. For the FZJ site we ascertained a value of 0.8. Again, for analysing the impact of different CUEs, we iterated over different CUEs at the FZJ site.

The test results in Figure 15 show that the Juggle utilization is again constant with different CUEs since the Optimizer schedules only some workload jobs to Juggle which can only run on that machine. Apart from these jobs no other ones were assigned to that cluster since its basic energy consumption is too high compared to the other testbed machines. Jufit shows a slightly higher utilization than Dune when the CUE value is on a similar level than the Dune one. However, this effect is reversed when the CUE value of FZJ site is increased. The Optimizer calculates then a higher emission on Jufit and generates therefore a higher utilization on Dune for meeting the CO_2 emission objective.

The runtime of a job on a dedicated machine can have an important impact on the overall energy consumption. Therefore, the plug-in's application benchmark feature was introduced to map the different performance of supercomputers when running jobs with known HPC

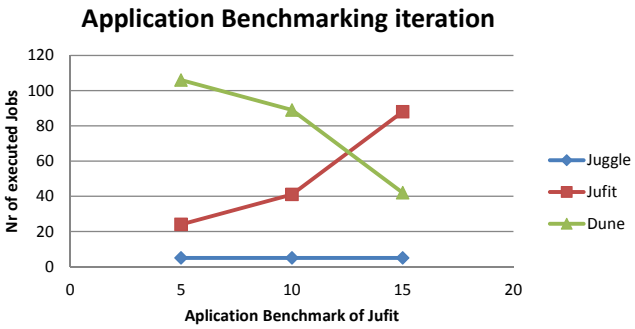


Figure 16. Impact of application benchmarking on job scheduling

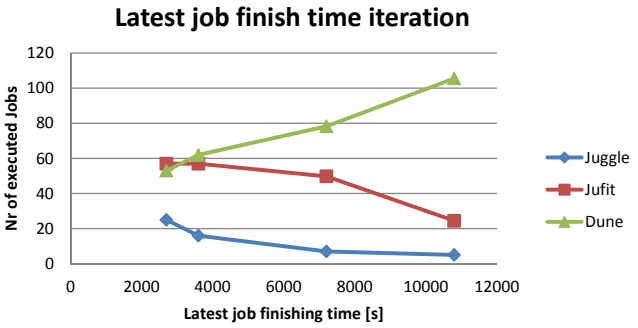


Figure 17. Job distribution and latest job finishing time

applications. The mapping is implemented by adding an application ID and a corresponding integer value, which indicates how efficiently the application can be executed.

In the test trials we used the LINPACK application for the evaluation of that parameter. LINPACK is installed on all available testbed clusters. Test benchmarks showed that all machines provide a different performance when executing jobs using the same LINPACK configuration. Finally, after measuring the execution time of the test jobs we were able to define different benchmark values for all the clusters of the testbed. Nevertheless, we wanted to evaluate not only this benchmark distribution but the impact of the application benchmarking on the job distribution and system utilization of the testbed. So, we performed additional tests with changing the benchmark parameter of the Jufit cluster and measuring the number of executed jobs on the clusters when iterating over different parameters (5, 10, and 15).

Figure 16 shows the distribution of the jobs during these different iterations. The Juggle cluster receives only a few jobs for all iterations. More revealing results could be detected in that evaluation on Jufit and Dune. When having similar benchmarks (Jufit 5, Dune 4) Dune receives clearly most of the workload jobs, since there is no big difference in that parameter between both machines and Dune is preferred by the Optimizer because of the node's better power consumption. When increasing the Jufit benchmark parameter from 5 to 15 it can be detected that Jufit gets as more jobs from the Optimizer as higher the benchmark parameter has been set. Finally, a benchmark value at Jufit of 3 times higher than the Dune value results that the actually more power efficient Dune machine is overpowered by the much better benchmark efficiency of the Jufit machine, i.e., even if Jufit consumes more power, it consumes less energy since it is expected to execute the job much faster than Dune. This shows that the benchmark application parameter can have a strong impact on the job distribution, which again affects the energy consumption of the federated clusters.

The Optimizer checks additionally if the user defined 'latest job finishing time' can be satisfied. This means that it is checked if the job can be executed and finished on a cluster within a user defined limit. If not the job cannot be scheduled in the most energy efficient way and the next cluster is chosen where the estimated wait time is smaller than the user defined value. By this mechanism the user can set a threshold value as part of the given job description from where on jobs must not be scheduled energy efficiently anymore. It is assumed that users will set a multiple value of the wall time of a job when considering the latest job finishing time. For example, if a job gets a walltime of 8 hours, it can be assumed that the user expects in general some wait time before his job can be executed. So, we analysed the impact of the latest finishing parameters which were set between three and ten times as high as the average walltime of the submitted workload jobs. Figure 17 shows the results of the job distribution when iterating the same workload with different job finishing parameters.

When relatively short latest job finishing times were set by users the Optimizer is not able to schedule most of the jobs in an energy efficient way. Taking into account the power estimation the Optimizer submits the jobs at first to Dune, calculates then that the overall finishing time (wait time + runtime) of the other waiting jobs would exceed their latest finish time, and thus will submit new jobs to other clusters until Dune provides again time slots for waiting jobs. In that way, it can be detected that even the Juggle cluster, which is usually underutilized by the federated plug-in scheduling algorithms because of its bad energy efficiency, gets more jobs as usual with low latest job finishing time parameters. With increasing latest job finishing times the plug-in has a greater margin to schedule jobs to energy efficient clusters, which is the Dune cluster at VTT in this evaluation.

In general, the latest job finishing time parameter is very important to guarantee as much as possible a fair sharing of the federated resources when there is high system utilization. On the other hand this parameter

helps to submit jobs each time to the most efficient cluster as long as the user's latest job finishing time can be satisfied. This condition can be rather satisfied in underutilized federated resources. Using additionally some of the developed single site algorithms to save energy on these underutilized resources would lead to further energy savings.

In the final assessment we analysed the overall energy and emission saving potential of the plug-in's federated algorithms. For this purpose, we used the constructed PUE and CUE values of the two testbed sites FZJ and VTT. While at VTT a CUE of 0.43 and a PUE of 1.2 has been defined, we set at Jufit and Juggle a CUE of 0.8 and a PUE of 1.4. For the LINPACK application benchmark parameter we used the results of our test measurements to detect the most efficient benchmark distribution, which is Juggle=1, Jufit=10, and Dune=4.

Concerning the latest job finishing time parameter we compared the energy consumption of different values since this parameter reflects very reasonably the overall utilization of the federated resources. The results were compared at first glance with the default round-robin strategy of the UNICORE middleware. However, the round-robin strategy is a bad algorithm in terms of energy efficiency. It just submits the jobs by considering time slots in the scheduling calculation and distributes thus the jobs pretty equally to the resources without regarding available idle resources, job queue information, and power consumption.

A much better algorithm in terms of scheduling jobs in federated environments without regarding the power consumption, PUE, and CUE values is the FP algorithm, which is a good reference point when comparing energy-aware- with non-energy-aware-algorithms in federated environments.

Table IX shows the results regarding energy consumption and produced emissions when using the plug-in for scheduling jobs efficiently in federated environments. The first column shows the measurement that was performed with a relatively low latest finishing time parameter of 3600 s while running test jobs with an average walltime of 720 seconds. Compared to the USO round-robin strategy one could save up to 52% energy when using the plug-in. When the latest finishing time is increased to 10800 s, this saving was even raised to 67%. This seems to be a very high saving. The reason is that the round-robin strategy submits in contrast to the plug-in a lot more jobs to Juggle. This results obviously in a much higher energy consumption of the Juggle nodes. Whereas more jobs on Jufit or Dune only raise the energy consumption moderately, on Juggle more jobs generate a significantly higher consumption.

The result of a measurement with the plug-in's FP algorithm is listed in the third column. Also that strategy schedules some jobs to Juggle. However, in that case, the waiting time of the queued jobs is checked continuously by the Optimizer, so that much less jobs run at the end on that machine. Moreover, considering the elapsed times of the workload on the particular clusters it can be detected that the total summarized elapsed time is shorter compared to the energy aware algorithm. Nevertheless, the FP algorithm is not as energy efficient as the energy aware strategy. At the end, the higher energy consumption of the most inefficient cluster is particularly disadvantageous for the total workload energy consumption. So, compared with the FP strategy and using same latest finishing times, the ECA algorithm saves about 21% energy. In the federated scenario, the energy-aware job scheduler (if used) causes the same 2-3 % overhead as in single site scenario. Additionally, FP and ECA algorithms cause some more overhead compared to the simpler round-robin algorithm that does not need any dynamic information from the clusters. Clearly, the overhead is much smaller than the achieved savings.

Table IX
Energy and Emissions in HPC federated measurements

Latest finish time [s]	ECA 3600	ECA 10800	FP 10800	RR -
Juggle [kJ]	4668	1916	3705	12893
CUE	0.80	0.80	0.80	0.80
PUE	1.40	1.40	-	-
Jobs	16	5	11	43
Elapsed Time [s]	2979	1245	2434	8181
Average Power [W]	1567	1540	1522	1576
Utilization [%]	66	56	46	81
Jufit [kJ]	961	767	1051	962
CUE	0.80	0.80	0.80	0.80
PUE	1.40	1.40	-	-
Jobs	57	25	61	45
Elapsed Time [s]	2940	2536	3213	2838
Average Power [W]	327	304	327	339
Utilization [%]	65	49	93	45
Dune [kJ]	1442	2163	1426	1029
CUE	0.43	0.43	0.43	0.43
PUE	1.20	1.20	-	-
Jobs	62	106	63	47
Elapsed Time [s]	3163	4721	3211	2839
Average Power [W]	456	458	444	430
Utilization [%]	82	97	88	79
Total cluster [kJ]	7072	4846	6181	14884
Saving to FP [%]	-	21.59	-	-
Saving to USO [%]	52.49	67.44	-	-
Total cluster emissions [kgCO2eq/kwh]	1.42	0.86	1.31	3.62

VI. Conclusion and Future Work

The results show that the generally used round-robin cluster selection algorithm can lead to unbalanced utilizations among clusters. This can be very inefficient in terms of energy consumption and CO_2 emissions. Using energy-aware job scheduling to power off idle computing nodes whenever possible greatly enhances the energy-efficiency. Load can also be balanced by replacing round-robin cluster selection by the Fastest possible selection algorithm. This leads to energy savings due to the better utilization of clusters and shorter wait times. Using both energy-aware job scheduling and FP cluster selection simultaneously leads to greater energy savings than using only one of them. The greatest CO_2 emission savings can be achieved by using ECA cluster selection algorithm to favour the cluster with least CO_2 emissions. The actual savings in each case depends on the cluster and job characteristics. In these simulations, for example, the energy sources were chosen so that one cluster had rather small CUE, another one rather big CUE, while the third one was something between them. With smaller differences in CUE, also the possible savings in CO_2 emissions would be smaller.

Based on the simulation results presented above, we propose to use FP cluster selection algorithm for the jobs without green SLA, since it leads to energy and CO_2 emission savings due to the better utilization of the clusters, and to better QoS due to the shorter wait time. For the jobs with green SLA, we propose to use the ECA cluster selection algorithm, since it can lead to even greater CO_2 emission savings than FP, while keeping the QoS (in terms of time) at the user specified level. It can be used also without green SLA, if some other parameter (e.g., queue size limit) is used for load balancing to prevent excessive load on the "greenest" cluster.

The results of the single site testbed assessment, with a focus on a DVFS feature, confirmed the experiences of our previous work [2] to the effect that the energy saving potential when using energy-aware scheduling increases with decreasing system utilization. When having loads above 80%, the saving potential is very limited (<10%). On the other hand utilizations below 60% show more reasonable savings (10% to 34%). The highest possible energy saving on the Juggle cluster was 34.7%, which happens when the system is completely idle. That value is limited by the fact that a compute node that was set to ACPI state standby cannot save more than 35% energy. It remains to be stated that the higher the system utilization the less compute nodes can be set to an energy saving status.

For the federated testbed results, it must be stated that the energy saving potential is absolutely dependent on the available hardware. A slow and high-consumption cluster as Juggle in a federated supercomputing environment can have a major impact on the results. If we had used three similar testbed clusters regarding their energy consumption, it can be expected that the energy saving potential would have been far smaller. In that case also a round-robin strategy would have produced better results and the FP strategy in the same way. Nevertheless, the performed test studies revealed many new insights into the saving potential of scheduling jobs energy efficiently in federated supercomputing environments. The results obtained can be used as a basis for the design of specific federated cluster environments when using the developed plug-in to enable an energy-aware scheduling of the resources.

The simulation studies and testbed experiments were performed with a different set of parameters e.g., hardware, type of jobs and cluster configurations. However, it is possible to note from both results that there is an energy saving potential in federated HPC environments that is dependent on the available hardware and cluster utilization. A poor utilization of clusters can ultimately lead to an increase in energy and emissions.

Previous research in the energy-efficiency of HPC grid computing has mainly focused on performing optimizations inside a single data centre. This work presented a global view by taking into account the whole grid: the characteristics of the data centres, compute nodes and the computing hardware. The most comparable approach to our work is HAMA, described in [14]. The results of HAMA are similar to our approach: energy savings are between 23% and 50%.

Recently, the plug-in's energy-aware job scheduler has been enhanced by adding several additional scheduling policies that are available in commercial schedulers like Moab. The new supported policies are mainly for enhancing the QoS for the users, including for example fair share and job exclusive policies. Fair share policy prevents a single user from conquering all the nodes with multiple jobs if there are also other users' jobs waiting in the queue. Job exclusive policy means that the job will get all the resources of the entire node for itself; only one job per node is then allowed. Our future work include testing the effect of these policies to the energy and emission savings.

Acknowledgment

This work extends our earlier work [1]; the invitation to submit this extended version is highly appreciated. This work was supported by EU FP7 project FIT4Green [33]. The authors would like to thank all the colleagues that have worked in the project, as well as the anonymous reviewers for their comments.

References

[1] M. Majanen and O. Mämmelä, "Optimization of Energy and Emissions in High-Performance Grid Computing Data Centres," in *The Second International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY 2012)*, St. Maarten, Netherlands Antilles, Mar. 2012.

[2] O. Mämmelä, M. Majanen, R. Basmadjian, H. Meer, A. Giesler, and W. Homberg, "Energy-aware job scheduler for high-performance computing," *Computer Science - Research and Development*, vol. 27, pp. 265–275, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00450-011-0189-6

[3] Y. Liu and H. Zhu, "A survey of the research on power management techniques for high-performance systems," *Softw. Pract. Exper.*, vol. 40, pp. 943–964, October 2010.

[4] Gartner. (2012, Sep.) Gartner Estimates ICT Industry Accounts for 2 Percent of Global CO_2 Emissions. [Online]. Available: http://www.gartner.com/it/page.jsp?id=503867

[5] The Climate Group, "SMART 2020: Enabling the low carbon economy in the information age," Tech. Rep., 2008.

[6] D. Erwin and D. Snelling, "UNICORE: A Grid Computing Environment," in *Euro-Par 2001 Parallel Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2001, vol. 2150, pp. 825–834.

[7] C. Belady, "Green Grid Data Center Power Efficiency Metrics: PUE and DCIE," Green Grid, Tech. Rep., 2008.

[8] M. Di Girolamo and et. al., "Final evaluation report on pilots of full-featured enhanced control plug-in and control desk for federated data centre," FIT4Green, Deliverable D6.4 v2.0, Jul. 2012, http://www.fit4green.eu/.

[9] S. Hong and H. Kim, "An integrated gpu power and performance model," in *Proceedings of the 37th annual international symposium on Computer architecture*, ser. ISCA '10. New York, NY, USA: ACM, 2010, pp. 280–289. [Online]. Available: http://doi.acm.org/10.1145/1815961.1815998

[10] D. Li, S. Byna, and S. Chakradhar, "Energy-aware workload consolidation on gpu," in *2011 40th International Conference on Parallel Processing Workshops (ICPPW)*, Sep. 2011, pp. 389 –398.

[11] K. Ma, X. Li, W. Chen, C. Zhang, and X. Wang, "Greengpu: A holistic approach to energy efficiency in gpu-cpu heterogeneous architectures," in *2012 41st International Conference on Parallel Processing (ICPP)*, Sep. 2012, pp. 48 –57.

[12] Moab Green Computing. (2012, Dec.). [Online]. Available: http://www.adaptivecomputing.com/resources/docs/mwm/archive/6-0/18.0greencomputing.php

[13] F. Ehmke, "Moab evaluation," University of Hamburg, Project Report, Dec. 2011, http://wr.informatik.uni-hamburg.de/_media/research/labs/2011/2011-09-florian_ehmke-moab_evaluation-report.pdf.

[14] S. K. Garg and R. Buya, "Exploiting Heterogeneity in Grid Computing for Energy-Efficient Resource Allocation," *Seventeenth Annual International Conference on Advanced Computing and Communications (ADCOM 2009)*, 2009.

[15] T. Lynar, R. Herbert, Simon, and W. Chivers, "Reducing grid energy consumption through choice of resource allocation method," *2010 International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW)*, May 2010.

[16] C. Patel, R. Sharma, C. Bash, and S. Graupner, "Energy aware grid: Global workload placement based on energy efficiency," Hewlett Packard, HP Laboratiories Palo Alto, Tech. Rep., November 2002.

[17] A. J. Shah and N. Krishnan, "Optimization of global data center thermal management workload for minimal environmental and economic burden," *IEEE Transactions on Components and Packaging Technologies*, vol. 31, no. 1, pp. 39–45, March 2011.

[18] G. Da Costa, J.-P. Gelas, Y. Georgiou, L. Lefevre, A.-C. Orgerie, J.-M. Pierson, O. Richard, and K. Sharma, "The GREEN-NET Framework: Energy Efficiency in Large Scale Distributed Systems," *HPPAC 2009 : High Performance Power Aware Computing Workshop in conjunction with IPDPS 2009*, May 2009.

[19] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Geographical load balancing with renewables," *SIGMETRICS Perform. Eval. Rev.*, vol. 39, no. 3, pp. 62–66, Dec. 2011. [Online]. Available: http://doi.acm.org/10.1145/2160803.2160862

[20] K. Le, R. Bianchini, T. D. Nguyen, O. Bilgir, and M. Martonosi, "Capping the brown energy consumption of internet services at low cost," in *Proceedings of the International Conference on Green Computing*, ser. GREENCOMP '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 3–14. [Online]. Available: http://dx.doi.org/10.1109/GREENCOMP.2010.5598305

[21] I. Goiri, K. Le, T. D. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenhadoop: leveraging green energy in data-processing frameworks," in *Proceedings of the 7th ACM european conference on Computer Systems*, ser. EuroSys '12. New York, NY, USA: ACM, 2012, pp. 57–70. [Online]. Available: http://doi.acm.org/10.1145/2168836.2168843

[22] Z. Liu, M. Lin, A. Wierman, S. H. Low, and L. L. Andrew, "Greening geographical load balancing," in *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, ser. SIGMETRICS '11. New York, NY, USA: ACM, 2011, pp. 233–244. [Online]. Available: http://doi.acm.org/10.1145/1993744.1993767

[23] Z. Liu, Y. Chen, C. Bash, A. Wierman, D. Gmach, Z. Wang, M. Marwah, and C. Hyser, "Renewable and cooling aware workload management for sustainable data centers," in *Proceedings of the 12th ACM SIGMETRICS/PERFORMANCE joint international conference on Measurement and Modeling of Computer Systems*, ser. SIGMETRICS '12. New York, NY, USA: ACM, 2012, pp. 175–186. [Online]. Available: http://doi.acm.org/10.1145/2254756.2254779

[24] I. Goiri, K. Le, M. Haque, R. Beauchea, T. Nguyen, J. Guitart, J. Torres, and R. Bianchini, "Greenslot: Scheduling energy consumption in green datacenters," in *2011 International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, Nov. 2011, pp. 1 –11.

[25] S. Klingert, T. Schulze, and C. Bunse, "GreenSLAs for the eco-efficient management of data centres," in *2nd International Conference on Energy-Efficient Computing and Networking 2011 (E-energy 2011)*, New York, NY, USA, 2011.

[26] C. Belady, D. Azevedo, M. Patterson, J. Pouchet, and R. Tipley, "Carbon Usage Effectiveness (CUE): A Green Grid Data Center Sustainability Metric," Green Grid, Tech. Rep., December 2010.

[27] RealtimeCarbon.org. (2012, Sep.) CO_2 conversion factors. [Online]. Available: http://www.realtimecarbon.org/resources/RealtimeCarbonMethodology.pdf

[28] W. Cirne and F. Berman, "A comprehensive model of the supercomputer workload," in *IEEE International Workshop on Workload Characterization, WWC-4. 2001*, dec. 2001, pp. 140–148.

[29] C. Bailey Lee, Y. Schwartzman, J. Hardy, and A. Snavely, "Are user runtime estimates inherently inaccurate?" in *Job Scheduling Strategies for Parallel Processing*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2005, vol. 3277, pp. 253–263.

[30] R. Basmadjian, N. Ali, F. Niedermeier, H. De Meer, and G. Giuliani, "A methodology to predict the power consumption of servers in data centres," in *Proc. of the ACM SIGCOMM 2nd Int'l Conf. on Energy-Efficient Computing and Networking (e-Energy 2011)*. ACM, 2011.

[31] OMNeT++. (2012, Sep.). [Online]. Available: http://www.omnetpp.org

[32] INET Framework. (2012, Sep.). [Online]. Available: http://inet.omnetpp.org/

[33] FIT4Green project. (2012, Sep.) FIT4Green: Energy aware ICT optimization policies. [Online]. Available: http://www.fit4green.eu

[34] V. Georgiadou, A. Salden, C. Dupont, A. Somov, A. Giesler, J. C. L. Egea, P. Barone, G. Giuliani, O. Abdelrahman, R. Lent, M. Kessel, T. Schulze, R. Basmadjian, H. D. Meer, M. Majanen, and O. Mämmelä, "Enhanced control plug-in and control desk software design specifications," FIT4Green, Deliverable D-5.3 v1.0, May 2012, http://www.fit4green.eu/.

[35] FIT4Green plug-in source code. (2012, Sep.). [Online]. Available: https://github.com/fit4green/FIT4Green

[36] Linux/ACPI - Documentation. (2012, Sep.). [Online]. Available: http://acpi.sourceforge.net/documentation/sleep.html

[37] LINPACK. (2012, Sep.). [Online]. Available: http://www.netlib.org/linpack/

[38] PEPC - Pretty Efficient Parallel Coulomb Solver. (2012, Sep.). [Online]. Available: www2.fz-juelich.de/zam/pepc

[39] Unicore Client Layer. (2012, Sep.). [Online]. Available: http://www.unicore.eu/unicore/architecture/client-layer.php

Towards the Live City – Paving the Way to Real-time Urbanism

Bernd Resch
SENSEable City Lab
MIT
Cambridge, US
berno@mit.edu

Inst. of Geography
University of Heidelberg, GER

Alexander Zipf
Inst. of Geography
University of Heidelberg, GER
alexander.zipf@geog.uni-heidelberg.de

Euro Beinat
Dept. of Geoinformatics
University of Salzburg, AUT
euro.beinat@sbg.ac.at

Philipp Breuss-Schneeweis
Wikitude GmbH
Salzburg, AUT
philipp.breuss@wikitude.com

Marc Boher
Urbiotica
Barcelona, ESP
marc.boher@urbiotica.com

***Abstract*—In contrast to projections, which stated that the wide-spread distribution of high-speed Internet connections would render geographical distance irrelevant, cities have recently gained importance in academic research. Yet, real-time monitoring of urban processes is widely unexplored. We present the concept of a *Live City*, in which the city is regarded as an actuated near real-time control system creating a feedback loop between the citizens, environmental monitoring systems, the city management and ubiquitous information services. After clarifying the term 'live' – as opposed to common understanding of 'real-time' – we identify four main barriers towards the implementation of the *Live City*: methodological issues, technical/technological problems, lacking quantification of economic revenues, and finally privacy and legislative questions. In this paper, we discuss those challenges and point out potential future research pathways towards the realisation of a *Live City* – ranging from sensor network developments, real-time quality assurance and new user interface paradigms to world-wide legislation measures, a standardised urban operating system and the idea of a 'tuned' city.**

Keywords-live city; ubiquitous sensor networks; real-time city; urban services; real-time information services.

I. INTRODUCTION

Based on the fast rise of digital communication technologies [1], projections stated that the wide-spread distribution of high-speed internet connections will render geographical distance irrelevant [2],[3], and that cities are not more than mere artefacts of the industrial age [4]. As a side effect, cities were presumed to drastically decrease in importance as physical and social connections, and would play an increasingly ancillary role in socio-technical research.

In reality, the world developed completely differently – cities are back in the centre of research. In fact, a United Nations (UN) report, which has been released before the World Population Day in 2007, states that for the first time in history, more people now live in cities than rural areas [5]. Thus, cities in their multi-layered complexity in terms of social interactions, living space provision, infrastructure development and other crucial human factors of everyday life have re-gained importance in scientific research. This arises from the fact – amongst others – that major developments of scientific and technological innovation took place in the urban context [6],[7].

However, in research on urban areas, especially real-time monitoring of urban processes and target-oriented deployment of digital services, are still widely unexplored. These research fields have recently received a lot of attention due to the fast rise of inexpensive pervasive sensor technologies, which made ubiquitous sensing feasible and enrich research on cities with uncharted up-to-date information layers through connecting the physical to the virtual world, as shown in Fig. 1.

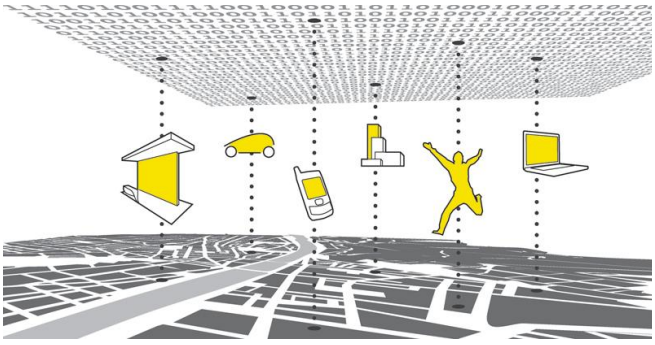


Figure 1. *Live City* – Connecting Physical and Virtual Worlds. [8]

One driver towards this vision is the diminishing digital divide on a global scale. While the digital divide within countries is still strongly affecting the degree of access to information and knowledge, the global digital divide is decreasing due to the fast rise of Information and Communications Technology (ICT) markets in China, India, South-East Asia, South America and Africa. Mobile phone penetration (mobile subscriptions per 100 inhabitants) has been at 76.2% of the world's population in 2010, where it is at 94.1% in the Americas and at 131.5% in Commonwealth of Independent States (CIS) [9]. The two fastest growing mobile phone markets China and India currently face a penetration rate of 64% and 70%, which makes a total number of 1.69 billion subscribers in those two countries alone.

This development builds the basis for the installation of urban real-time services. In a recent report on Digital Urban Renewal [10], the author states that major demand-side drivers for digital urban projects are the increasing focus on sustainability and emissions reduction, continued pressure on the urban transport infrastructure, and increasing pressures on citizen services due to demographic shifts, amongst others. On the supply side, several drivers have been identified including the ongoing evolution of the Internet as an underlying framework for services, new connectivity technologies, sensor networks and augmented reality.

All these comprehensive ideas supporting the paradigm of assessing, analysing and influencing urban environments in (near) real time (s. Section III for a disambiguation of 'real-time' and 'live') require a number of cognitive concepts, spatio-temporal algorithms and technological developments to be feasible.

However, we are still facing a lack of experience in assessing urban dynamics in real time. One reason is that ambient and continuous monitoring is an enormous challenge, and this is particularly true in the urban context, which poses very specific challenges. These comprise well-known technological questions, but also significant economical, social and political ones, which are rapidly gaining importance. This applies to a wide range of recent developments connected to live cities such as the Internet of things, pervasive sensing or ubiquitous urban monitoring.

Over the last few years, researchers and practitioners have also dealt with the apparent disconnect between the technical capabilities developed by researchers and technology firms in the broad context of smart cities and the actual adoption rates in cities. It can be argued that this follows the normal pattern of innovation adoption, and that in Geoffrey's Moore language the technologies have not yet "crossed the chasm" of adoption. However, a few patterns seem to emerge in urban environments, stressing the social and organisational nature innovation in addition to the technology aspects.

An important aspect to mention is that real-time and smart cities are nowadays often associated and developed under the umbrella of energy-related questions and applications, such as the Strategic Energy Technology initiative [11] by the European Commission. This paper tries to present a more holistic and comprehensive definition of a *Live City*. We see the city as a multi-layered construct containing multiple dimensions of social, technological and physical interconnections, i.e., as an actuated multi-dimensional conglomerate of heterogeneous processes, in which the citizens are the central component.

Towards the realisation of a *Live City*, we are currently experiencing a fast progressing technology development, which is not only moving ahead quickly, but which is moving ahead of society. This development can be compared with a stream moving at high speed, on which we are paddling to remain on the same spot or at least not to drift off too fast. The question, which we have to tackle in this regard, is where our goal for the future lies: down-stream, somewhere near our current spot, or even up-stream?

In this paper, we try to illustrate possible pathways to answering this multi-dimensional question. We incorporate societal, technical, political, privacy and economic issues into our rationale. We are well aware of shortcomings in terms of completeness and technical thoroughness. The paper shall be considered a first leap towards a *Live City* 'Installation Guide'.

This paper is organised as follows: after this introduction we illustrate a few examples on existing approaches towards *Live Cites* in Section II before giving a disambiguation of the term 'live' in Section III. Section IV discusses challenges in current research on the *Live City* and Section V illustrates potential future research avenues, before Section VI summarises conclusions from the paper.

II. STATE-OF-THE-ART – REAL-TIME AND LIVE CITIES

One of the first implementations of a 'real-time city' has been done by the MIT SENSEable City Lab [8]. This research group has considerably coined the term '**real-time city**', particularly through visualising the city as a real-time and pulsating entity. In further research initiatives, the SENSEable City Lab investigated human mobility patterns, the usage of pervasive sensors to assess urban dynamics, event-based anomaly detection in ICT infrastructures, and correlations between ICT usage and socio-cultural developments. The major shortcoming in this research is that no generic long-term goals are addressed apart from singular implementations in selected cities.

A new and innovative idea in the context of assessing urban dynamics in real time is the concept of **Living Labs**. According to [12], a Living Lab is a 'real-life test and experimentation environment where users and producers co-create innovations'. Living Labs are strongly driven by the European Commission, which characterises them as Public-Private-People Partnerships (PPPP) for user-driven open innovation. A Living Lab is basically composed of four main components: co-creation (co-design by users and producers), exploration (discovering emerging usages, behaviours and market opportunities), experimentation (implementing live scenarios) and evaluation (assessment of concepts, products and services). Even though the concept of Living Labs has rapidly gained attention over the last decade, it has not been holistically explored in terms of general research challenges and concrete future points of action.

Also, much research is performed in the area of **smart cities** (in particular in South Korea also the term 'ubiquitous cities' is popular). For instance, IBM has implemented a number of urban services in the course of their 'Smarter Planet' programme [13]. Within this initiative research is performed together with cities all over the world to implement applications in the areas of city management, citizen services, business opportunities, transport, water supply, communication and energy. The goal is to seize opportunities and build sustainable prosperity, by making cities 'smarter'. Despite the seminal nature and the broad awareness for the concept, smart cities are currently mostly understood in terms of energy-related questions (particularly in the European Union), i.e., nearly no trans-disciplinary approaches have been developed.

A sensor-driven approach to ubiquitous urban monitoring is presented in [14] and in [15]. The authors present a measurement infrastructure for **pervasive monitoring** applications using ubiquitous embedded sensing technologies with a focus on urban applications. The system has been conceived in such a modular way that the base platform can be used within a variety of sensor web application fields such as environmental monitoring, biometric parameter surveillance, critical infrastructure protection or energy network observation. Several show cases have been implemented and validated in the areas of urban air quality monitoring, public health, radiation safety, and exposure modelling. Yet, these approaches mainly focus on technical and methodological developments and do not account for wider challenges such as societal, political and legislative aspects.

III. A DISAMBIGUATION OF THE TERM 'LIVE'

The term '*Live City*' originates from the modification of the expression 'Real-time City' as definitions and usages of the latter expression are vague and vary on a quite broad scale.

Anthony Townsend presents a highly mobile phone centric definition of a real-time city by stating that 'the cellular telephone […] will undoubtedly lead to fundamental transformations in individuals' perceptions of self and the world, and consequently the way they collectively construct that world' [16]. The author sees the real-time city as a potential platform for dedicated advertising and states that 'accessibility becomes more important than mobility'. This implies that it will be more critical to access urban services rather than moving around physically. This in turn means that the digital (i.e., mobile phone) infrastructure will be more important than the physical (i.e., transport) infrastructure.

A possible definition of *urban informatics* – a term closely related the real-time city – is 'the collection, classification, storage, retrieval, and dissemination of recorded knowledge of, relating to, characteristic of, or constituting a city' [17]. This definition gives a more holistic, but rather general view on the term 'real-time city', which centers around information and knowledge while cultural, social, political and privacy aspects remain greatly untouched.

Apart from these definitions, the term is generally understood as providing spatial information about the city in a timely manner without necessarily accounting for a feedback loop or dynamic processes.

In these interpretations of the expression 'real-time', it has been strongly mitigated. The term 'real-time' originated in the field of computer science, where it initially described a process, which is completed 'without any delay'. This broad view was then divided into hard and soft real-time demands. Soft real-time basically defines that deadlines are important, but the whole system will still function correctly if deadlines are occasionally missed. The latter is not true for hard real-time systems. Another term to express non-rigorous temporal requirements is 'near real-time', which describes a delay introduced into real-time applications, e.g., by automated data processing or data transmission [18]. Hence, the term accounts for the delay between the occurrence of an event and the subsequent use of the processed data.

These definitions of the term 'real-time' have been set up for the domain of computer science. Thus, it is important to evaluate and re-define the expression in the context of urban geography. Naturally, strict real-time requirements are a central aspect in monitoring applications, whereby these demands are highly application-specific and can vary significantly. Therefore, they are not a fundamental goal in the field of urban geography, as the term 'real-time' is primarily defined by an 'exact point in time', which is the same for all data sources to create a significant measurement outcome. Secondarily, the term defines the possibility to start a synchronous communication at a certain time, which might often be important for geographical monitoring applications, e.g., to enable the generation of an exact development graph for temporal pollutant dispersion over a defined period of time in precise intervals.

Additionally to the suggestion of assessability of the environment in the 'now', the expression '*Live City*' also implies a feedback loop. The term 'city' does not only define the description of location-aware parameters, but also entails the exploration of causal patterns in these data. In the context of geo-sensor network and monitoring applications, this in turn means that the urban environment is not only analysed remotely by examining quasi-static data, but the procedure of sensing and processing live data offers the possibility of modifying the urban context in an ad-hoc fashion.

In conclusion, it can be stated that the strict term 'real-time' can be interpreted as 'at present' for urban monitoring applications, in the sense that the aim is to assess the environment 'now', not a historical and perhaps outdated representation. However, these topicality requirements can vary depending on the application context. For instance, an update on traffic conditions does not have to exceed a delay of a couple of minutes when this information is used for navigation instructions, whereas a 30 minute update interval can well be sufficient for short-term trip planning.

To account for this non-rigorous requirement, the term '*Live City*' seems better suited than 'Real-time City'. In this reflection, 'near real-time' appears to be closest to 'live', as it does not impose rigid deadlines and the expression itself suggests dynamic adaptation of a time period according to different usage contexts.

IV. CHALLENGES IN CURRENT RESEARCH ON THE LIVE CITY

The urban context poses many challenges to pervasive monitoring and sensing systems. Particular issues arise for the deployment of near real-time information services in the city. These range from physical sensor mounting to social and privacy implications. Furthermore, the sensitive urban political landscape has to be accounted for, which might cause unforeseen challenges. Naturally, technical challenges play a key role in the establishment of *Live Cities*.

A. Technological and Technical Issues

The first essential technological challenge is the integration of different data sources owned by governmental institutions, public bodies, energy providers and private sensor network operators. This problem can potentially be tackled with self-contained and well-conceived data encapsulation standards – independent of specific applications – and enforced by legal entities, as discussed in sub-chapter V.B. However, the adaptation of existing sensors to new interoperability standards is costly for data owners and network operators in the short term, and so increased awareness of the benefits of open standards is required.

From a technical viewpoint, unresolved research challenges for ubiquitous urban monitoring infrastructures are manifold. These challenges range from finding a uniform representation method for measurement values, optimising data routing algorithms in multi-hop networks, data fusion, and developing optimal data visualisation and presentation methods. The latter issue is an essential aspect in real-time decision support systems, as different user groups might need different views on the underlying information. For example, in case of emergency local authorities might want a socio-economic picture of the affected areas, while first-response forces are interested in topography and people's current locations, and the public might want general information about the predicted development of a disaster.

In addition, there are a number of well-known technical issues in the establishment of urban monitoring systems (energy supply, sensor mote size, robustness, routing, ad-hoc network connections, reliability, connectivity, self-healing mechanisms, etc.). These have to be addressed as the case arises depending on specific requirements of the end application. Thus, they are not part of the presented research.

Furthermore, highly unpredictable challenges exist arising from the openly accessible, dynamic and variable urban environment, such as severe weather conditions, malfunctioning hardware, connectivity, or even theft and vandalism. These general and mostly technical issues are well elaborated and shall not be discussed within this paper.

B. Various Stakeholders

Other issues for the installation of a *Live City* are thematic challenges and socio-political concerns, which are rapidly gaining importance. The feedback loop depicted in Fig. 2 is a key factor in designing urban monitoring systems. In practice, various kinds of stakeholders have to be considered including citizens, information providers, research institutions, politicians, the city management, or other influential interest groups. This cycle involves all steps of the deployment process from planning, deployment, customised information provision, and feedback from the citizens and other interest groups [19].

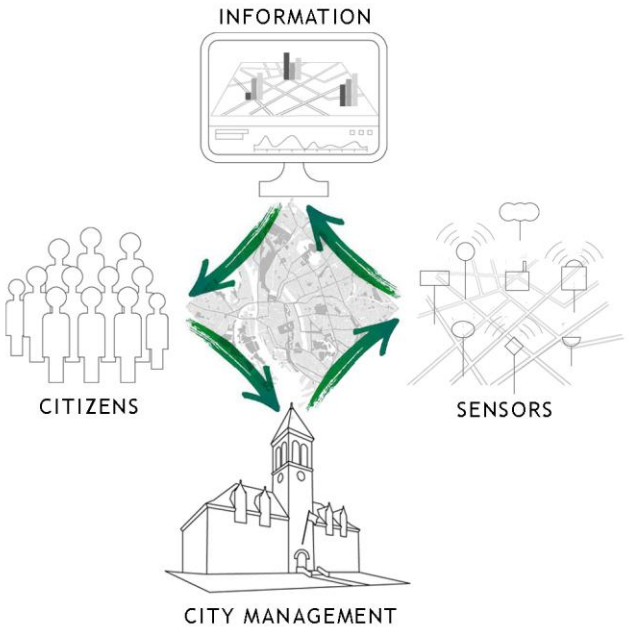


Figure 2. Feedback Loop Enabling the *Live City*.

Regarding the cycle depicted in Fig. 2, a common pattern refers to the network of actors that constitute the city from the decision-making point of view. While for simplicity we may indicate "city management" as the entity that supervises investments and decisions, in practice the end result of city innovations is the consequence of decisions made by a network of public and private actors, each one with a separate agenda and specific political or commercial objectives. The synchronisation, or lack thereof, between these actors frequently determines speed, depth and extent of adoption of new technologies, and the impact that technology ultimately has on the city. The process of public-private partnerships (PPP) and the governance of technology innovation in urban environments are frequently secondary concerns of technology vendors and scientists.

Another important methodological peculiarity of the urban context is that there are large variations within continuous physical phenomena over small spatial and temporal scales. For instance, due to topographical, physical or radiometric irregularities, pollutant concentration can differ considerably, even on opposite sides of the street. This variability tends to make individual point measurements less likely to be representative of the system as a whole. The consequence of this dilemma is an evolving argument for environmental regulations based on comprehensive

monitoring data rather than mathematical modelling, and this demand is likely to grow.

C. The Value of Sensing Collective Behaviour versus Privacy Implications

Although we experience quickly increasing awareness of the opportunities of digital mobile communication, the question arises how we can engage people to contribute actively being 'human data sources' and involve themselves into re-designing urban processes. This is necessary in order to leverage collective information in areas such as environmental monitoring, emergency management, traffic monitoring, or e-tourism. One example, in which this kind of volunteered data was of invaluable importance, were the earthquake and the subsequent tsunami in Japan in March 2011. In this case, the *Tweet-o-Meter* [20] application has been used to find anomalies in Twitter activity. Right after the earthquake, people started to post status reports, video streams, and conditions of destroyed houses and cities, which could be interpreted in near real time as an indicator for an extraordinary event. Furthermore, information could be semantically extracted from personal comments and posts. In this context, an important but yet poorly research issue is the use of incentive schemes to encourage people to contribute their data. Current approaches mostly comprise 'feedback' and 'gamification', but their practical suitability has not been fully proven yet.

Early implementations and deployments also show the impact of new sources of evidence on the way organisations operate. As an example, police forces can greatly benefit from real-time feeds of urban information showing anomalies in the collective city behaviour that may underline safety risks or attention areas. While this is increasingly feasible through data mining of real-time data feeds, the challenge for law enforcement is how to react to this information, which questions the normal planning and operational practices.

Faced with real-time or even predictive safety alerts, should the police forces adopt a dynamic allocation of resources based on demand and supply mediated by technology or use this input as an extra source of evidence for existing management practices? How does this impact labour, skills, equipment or work shift? What are the legal or institutional implications? All city organisations are faced with similar questions, which emerge because technologies have enabled new information streams forcing them to reconsider established practices and operations. While it can be argued that this is the case for any organisation facing disrupting technologies, in urban settings the issue is compounded by the fact that many institutions face the same challenge simultaneously, and the degree to which organisations succeed or fail depends only in part on themselves.

This development raises the challenge to find the balance between providing pervasive real-time information while still preserving people's privacy. Strategies to address this stress field are described in sub-Section 0. In addition, it seems self-evident that the provided information has to be highly accurate, reliable and unambiguous. Thus, quality control and error prevention mechanisms including appropriate external calibration are even more important for monitoring networks in the city than in other, less connected, environments. The issue of quality control will be further discussed in sub-Section V.A.

In terms of privacy, the claim might arise that we need to be aware of our personal and private data *before* we share them. The essential question in this context, however, is *how* we can raise awareness of ways to deal with that matter. Terms and conditions of digital services and technology are mostly hardly understandable to non tech-experts. Thus, more simple and binding ways of communicating this kind of information have to be found.

Finally, some more unpredictable challenges posed by the dynamic and volatile physical environment in the city are radical weather conditions, malfunctioning hardware, restricted connectivity, or even theft and vandalism. Moreover, there are a number of rather obvious but non-trivial challenges to be addressed, such as optimal positioning of sensors, high spatial and temporal variability of measured parameters or rapid changes in the urban structure, which might cause considerable bias in the measurements.

V. DISCUSSION: FUTURE RESEARCH AVENUES

From the challenges described in Section IV we can derive a number of essential research questions, which have to be tackled in the area of *Live Cities*. These can be divided into methodological aspects, technical and technological issues, questions on privacy and legislation, and the assessment of economic benefits, which arise through the installation of a *Live City*.

A. Methodological Research

Over the last years prospects were made that 'data would be the new oil' [21],[22]. It has been stated that - like oil - data cannot be used without first being refined. This means that raw data is just the basic ingredient for the final product of **contextual information** that can be used to support strategic and operational decisions. Thus, a central issue in terms of providing real-time information services is the analysis of data according to algorithmic requirements, representation of information on different scales, context-supported data processing, and user-tailored information provision aligned with the needs of different user groups.

In general, the deployment of a large number of sensors ensures more representative results together with an understanding of temporal and spatial variability. However, deploying sensor networks is costly, politically sensitive and requires much time. One way to overcome these issues is to 'sense people' and their immediate surroundings using everyday devices such as mobile phones or digital cameras, as proposed by Goodchild [23]. These can replace – or at least complement – the extensive deployment of specialised city-wide sensor networks. The basic trade-off of this people-centric approach is between cost efficiency and real-time fidelity. The idea of using **existing devices to sense the city** is crucial, but it requires more research on sensing

accuracy, data accessibility and privacy, location precision, and interoperability in terms of data and exchange formats.

In terms of geo-data sources, Volunteered Geographic Information (VGI) plays a key role in realising the idea of a *Live City*. We are already experiencing an overwhelming willingness of citizens to contribute their personal observations ranging from opinions posted on Facebook to Tweets about local events or commented photo uploads on Flickr. As mentioned in Section IV, this kind of **collective information** can potentially have a vital impact on operational real-time strategies in areas such as emergency management, dynamic traffic control or city management.

A central issue in VGI is the **representativeness of volunteered information** [23],[24]. We argue that defining or deriving consistent semantics in user-generated content possibly requires the combination of bottom-up and top-down approaches. In bottom-up approaches, user communities build their own semantic objects and connections between those by using their own personal taxonomies. In contrast, top-down approaches – mostly academically driven – try to define semantic rules and ontological connections in a generic way prior to and independently of the end application.

Only the combination of those using Linked Data concepts (rather than rigid and inflexible ontology approaches) can lead to **domain-independent and comprehensive semantic models**, which are needed to cover the whole breadth of topics, users and applications in the *Live City*. This requires standardisation on two levels – firstly on sensor data level (encodings for measurements) and secondly on phenomenon level (measurand encodings). In this regards, semantic search will be an essential concept to extract knowledge and information from user-generated data combined with sensor measurements.

An aspect, which is strongly connected to availability of data sources, is **openness of data**. As argued by Jonathan Raper [25], quality of decision-support is increasing with the quality and the quantity of available data sources. We are currently facing a situation that in most cases, too little data are available to support well-informed decisions in near real time. This raises the question how data owners such as companies in the environmental sector, energy providers or sensor network operators can be animated to open their data repositories for public use.

On the contrary, we might face a vast amount of data freely available in the near future, contributed by a variety of different data producers – mostly non-quality assured data stemming from private observations or sensor networks. This of course raises the question of trustworthiness of these data. Thus, **automated quality assurance** mechanisms have to be developed for uncertainty estimation, dynamic error detection, correction and prevention. In this research area,

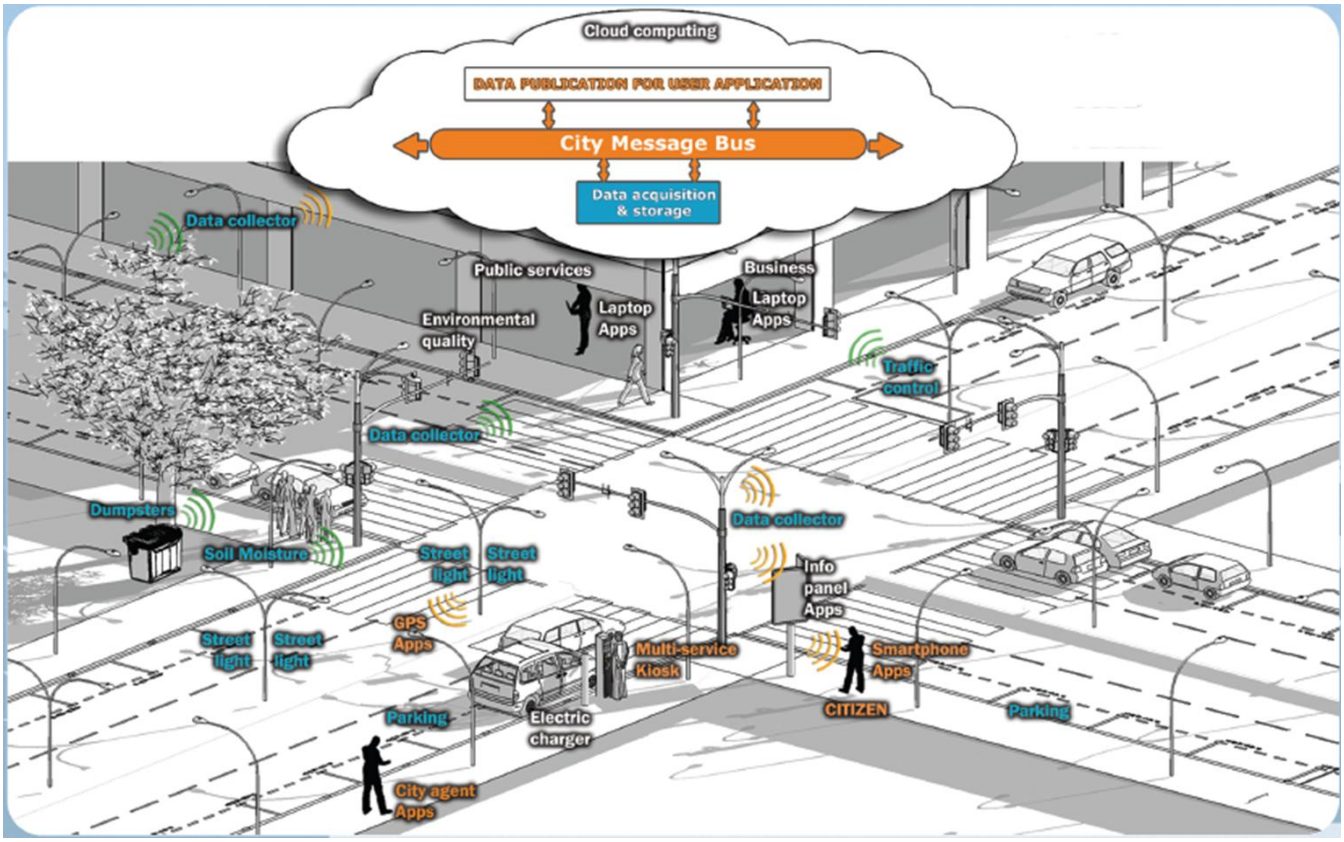


Figure 3. Urban Operating System.

we are currently seeing different approaches in development including Complex Event Processing (CEP) [14] for error detection, standardisation efforts for representing uncertainty in sensor data (e.g., Uncertainty Markup Language - UncertML) [26], or proprietary profiles to define validity ranges for particular observations. Only when these questions are solved, reliability and completeness of recommendations can be ensured.

Up to now quality analysis on VGI (e.g., [27],[28],[29]) focused on completeness as well as geometrical and semantic accuracy, but considered only few timeslots and neglected the (near) real time aspect of a *Live City*. Interestingly this kind of data can also be used to improve the data situation through intelligent algorithms as [30] shows.

Furthermore, measurements are only available in a quasi-continuous distribution due to the high **spatial and temporal variability** of ad-hoc data collection. Addressing this issue will require complex distribution models and efficient resource discovery mechanisms in order to ensure adaptability to rapidly changing conditions.

B. Technical and Technological Research

Practical experiences show that urban growth necessitates management with focus on efficiency and durability. The careful integration and managing of acquired information and ensuring a bigger and better interoperability of various services will ultimately drive successful management of the urban space. Consequently cities need an '**urban operating system**' which will endow them with new intelligence in coordinating and interconnecting all services. Those services comprise provision of real-time information on mobile devices facilitating smarter movement around the city, and the optimisation of services for city and contractor agents of public service to improve information exchange and to provide public access to open data in order to encourage citizen participation.

The urban operating system, illustrated in Fig. 3, consists of a setup of material and software architecture and allows for addressing the challenges mentioned in Section IV by taking into account constraints linked to outdated infrastructure:

- Acquisition of the information in real time
- Transportation of the information from public road network to information system
- Integration of systems deployed in the city (parking spaces, streetlights, traffic systems, waste management, etc.)
- Processing, dissemination, publication and storage of information in real time (real-time publication and dissemination, and provision of historical data for dynamic data mining processes)

In effect, the urban operating system acts as a connecting base layer, which enables the interplay of urban objects, public infrastructure and the citizens. Like this, people can benefit by gaining access to real-time information about the city (traffic conditions, air quality, social activities, public transport, health-related issues, etc.). This entails citizens to base their short-term decisions on real-world conditions, which are conveyed on demand in near real time.

This in turn requires the continuous assessment of urban processes, which requires the broad installation of sensor networks. The deployment of sensor networks implies a number of challenges, particularly in urban settings. Apart from technical research in the area of sensor networks regarding miniaturisation, energy supply, robustness, ad-hoc network connections, reliability, connectivity, self-healing mechanisms, etc., **standardisation and interoperability** are vital prerequisites for establishing pervasive and holistic monitoring systems. As current sensor network implementations are mostly built up in proprietary single-purpose systems, efforts to develop a uniform communication protocol will be needed [31]. One very promising approach in this field is the Sensor Web Enablement (SWE) initiative [32] by the Open Geospatial Consortium (OGC). SWE aims to make sensors discoverable, accessible and controllable over the Internet. SWE currently consists of seven standards and interoperability reports, including the Sensor Observation Service (SOS) for observation data retrieval, Observations and Measurements (O&M) for sensor data encoding, Sensor Markup Language (SensorML) for platform description and the Sensor Alert Service (SAS) for event-based data transmission. More details about SWE can be found on the

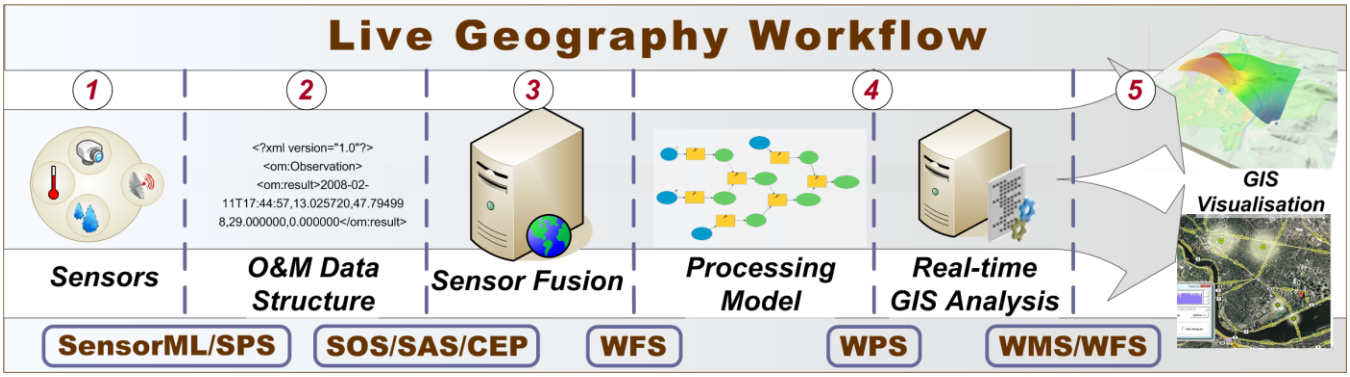


Figure 4. Live Geography – Standardised Geo-Sensor Data Analysis Architecture.

OGC web site[1].

In terms of integrating various kinds of real-time data such as sensor measurements, meteorological data, energy system states and human observations with existing sensor and analysis systems, the creation of a **standardised measurement infrastructure** using well-conceived data and service standards is a major technical challenge. An essential factor will be the integration of new developments with existing community projects such as Ushahidi or Twitter to create new possibilities for high-quality information. This will potentially result in increased situational awareness and enhanced Common Operational Pictures (COP). Furthermore, this development supports infrastructure-oriented approaches and directives such as Global Monitoring for Environment and Security (GMES), Infrastructure for Spatial Information in Europe (INSPIRE) and the Shared Environmental Information Space (SEIS). The *Live Geography* infrastructure [31] is a first step towards the realisation of a generic workflow-oriented geo-sensor data analysis architecture. As shown in Fig. 4, the Live Geography workflow covers all essential parts including standardised sensors, sensor fusion, web-based data processing and geo-data visualisation, and is generic enough to integrate all data sources relevant to Live City applications.

Here, an important future research area is the development of generic and portable **sensor fusion** algorithms, which are a vital prerequisite to combine data stemming from different heterogeneous sensor networks. Sensor fusion basically stands for the harmonisation of data in terms of units of measure, time zones, measurement models and observation semantics. To be compliant with the requirements of a *'Live' City*, the fusion process has to happen in near real time. [33] presents an approach for on-the-fly integration of measurements coming from different SOS instances using the free open-source server GeoServer[2]. The system harmonises measurements in real time and provides them on the fly via standardised OGC web service interfaces such as the Web Feature Service (WFS) and the Web Map Service (WMS). Although this implementation is still improvable in terms of fusion capabilities, it demonstrates a seminal approach towards sensor fusion.

The next step in a *Live City* workflow is geo-analysis of real-time data sources, i.e., the **refinement of raw data** towards user group-specific information, which can then be used for short-term decision support. This analysis process can be implemented by the OGC Web Processing Service (WPS) in a standardised way in general. But the WPS architecture is very generic in its current version so that the developments of further specialised (domain-specific) application profiles are necessary as is argued in [34],[35] and [36]. The power of using WPS for implementing more complex analysis functionality for urban models has for instance been shown in [37].

Another methodological issue in terms of communication technology is the availability of **ubiquitous communication media**. Today, we presume a fully functioning Internet to transmit information. However, in case of emergency, this layer is potentially not available, as we experienced for instance during hurricane Katrina in 2005 in New Orleans. Thus, we have to find alternate possibilities to communicate critical information independently of existing infrastructures. Possible solutions comprise long-range ad-hoc networks or the construction of a robust communication core network, which can withstand external influences such as tsunamis, earthquakes, storms, avalanches or even vandalism.

The *Live City* concept naturally implies the **provision of user-tailored information in near real time**. However, we have to consider that fulfilling this requirement is not always possible, for instance if data analysis algorithms are very complex and laborious, or if base data are only updated in certain intervals. Despite these restrictions we have to find algorithmic methods to accurately predict developments in our environment even in case of reduced data availability. This can for instance happen through the integration of well-calibrated models and spatio-temporal interpolation algorithms. This approach can naturally only mitigate the drawback of imprecise information, but not eliminate it.

A tightly related research challenge is the creation of **ubiquitous user interfaces**. Here, we tend to think about visionary devices and gadgets, but even today we have the need to develop solutions for some application domains like *Live Cities*, where an extremely wide range of legacy user interfaces is being used at different places, but still requiring a stringent link amongst each other. This means that people need to share the same information and be aware of other users of the system – in particular their location. The roles of these users are highly dynamic and change with time and space – so do the user interfaces they need to interact with up-to-date shared knowledge and each other.

A wide range of users exists within the *Live City* – from the general public, management centres, governmental institutions and urban planners to researchers and specialised maintenance staff. All of them need to access and interact with geospatial information in an truly ubiquitous way, i.e., simultaneously at different places with different devices: in the management central on interactive large screen wall displays, outdoor with tablet computers or smartphones, or potentially with new **augmented reality** devices – in particular as hands need to stay free for more authentic and pure experiences of the city.

Even wearable computing appliances start to play a central role and have to be considered for special. In this context, a highly interesting aspect is that we do not only have alphanumeric data to be presented, but the users need to interact with highly **interactive spatiotemporal information** (2D+t and 3D+t) on this broad range of devices – but in all cases users need to access similar functionality and information in an ad-hoc fashion. Some of this information is public or shared, some is restricted to distinct users or user roles. Still, there is a need of consistency – at least on the mental level for the interaction and visualisation metaphors used. This consistency is required because people might get confused and make errors if they need to switch

[1] http://www.opengeospatial.org

[2] http://www.geoserver.org

too often between different paradigms – something that needs to be minimised extremely in real-time applications.

Another interface-related problem is that *Live City* applications can often not rely on the wireless network – as mentioned above. In consequence, there is a need for supporting both autonomous and **collaborative decision-making and interaction**. This makes user interface design for different user groups a particular challenge. Currently we have to deal with a wide range of these challenges in the case of city management systems using today's technology (also to guarantee robustness of the application), but it is clear that new ubiquitous UI devices, metaphors and paradigms have a large potential in this area.

This also includes the provision of **real-time data and information in 3D**, as illustrated in Fig. 5. Integrating up-to-date sensor data, such as environmental, traffic-related or safety-relevant data can significantly improve the value of urban information systems. This requires the visualisation of 3D city and landscape models and the interactive navigation through the scene, which again raises a number of research challenges such as representation of 3D data on small screens, optimised information reduction, guaranteeing highest-possible representativeness of the data, creating urban dispersion models or sufficient update cycles.

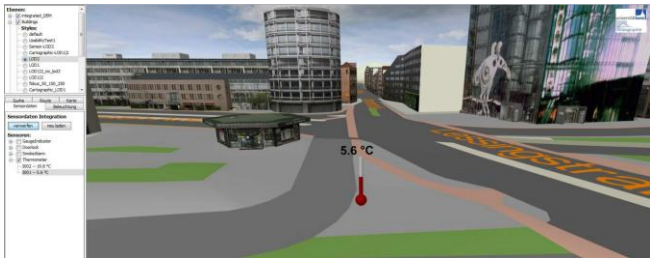


Figure 5. Integration of Sensor Data in 3D Environments. [38]

C. *Privacy and Legislation Measures*

Particularly in the area of user interaction and public participation procedures, a crucial question in the context of *Live Cities* is how we can **preserve people's privacy** dealing with ubiquitous information and partly personal data. One possible solution to address this issue is to make use of new Collective Sensing approaches. This methodology tries not to exploit a single person's measurements and data, but analyses aggregated anonymised data coming from collective networks, such as Twitter, Flickr or the mobile phone network [39]. Like this, we can gain a coarse picture of the situation in our environment without involving personal details of single persons. In case of tracking applications or services, in which personal data are involved, people have to have an opt-in/opt-out possibility. This means that users can decide themselves whether they want to use the application – and also withdraw their consent - being aware of the type and amount of data that is collected and transmitted.

Another central issue in deploying monitoring systems in the city is the personal impact of fine-grained urban sensing, as terms like 'air quality' or 'pollutant dispersion' are only surrogates for a much wider and more **direct influence** on people, such as life expectation, respiratory diseases or quality of life. This raises the demand of finding the right level of information provision. More accurate, finer-grained or more complete information might in many cases not necessarily be worthwhile having, as this could allow for drawing conclusions on a very small scale, in extreme cases even on the individual. This again could entail a dramatic impact in a very wide range of areas like health care, the insurance sector, housing markets or urban planning and management.

A central question in this context is: can we actually achieve a system, in which transactions are not tracked or traced? Thinking about mobile phone calls, credit card payments or automated toll collection, each of the underlying systems has to have some kind of logging functionality in order to file payments and generated automated reports. In these cases it is probably just not possible prevent storage – at least for a short time. Thus, **legal frameworks** have to be developed on national, trans-national and global levels. The largest limiting factor in this regard is the varying interpretation of 'privacy' in different parts of the world. For instance, privacy can be traded like a good by its owner in the USA, whereas it is protected by law in the European Union. This means that supra-national legislation bodies and initiatives are called upon to set up appropriate world-wide regulations.

As shown in Fig. 6, legislation and governments play a highly different role in these two settings.

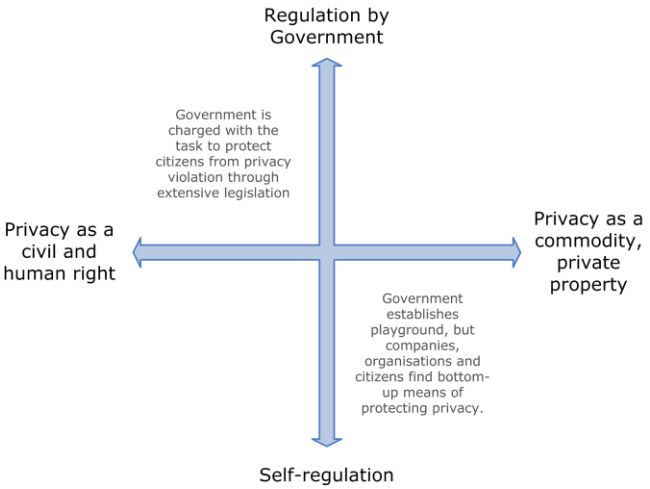


Figure 6. Different Understandings of Privacy.

This also includes the critical question of **data ownership** – who owns the data: the data producers (i.e., the citizens or a mobile phone network operator), the institutions that host a system to collect data, or the data providers? Furthermore, if sensitive data is analysed to produce

anonymised information layers, who is responsible if decisions that are based on this information are wrong due to lacking quality of the base data? In conclusion, the issues of privacy, data ownership, accessibility, integrity and liability have to be tackled thoroughly all at once and not separately from each other.

In case of tracking applications or services, in which personal data are involved, people should have an **opt-in/opt-out** possibility. This means that users can decide themselves whether they want to use the application – and also withdraw their consent – being aware of the type and amount of data that is collected and transmitted.

D. Assessing the Economic Value of Live Cities Services and Applications

Finally, an important aspect is the assessment of the economic value of establishing a *Live City*. Concrete revenues have not yet been defined, which would compel this kind of investment. Thus, we need to find instruments to quantify financial benefits of ubiquitous information services, city-wide sensor networks and mobile applications collecting user-generated information on the current status of the city.

It is symptomatic that previous experiences demonstrate that the understanding of the network of decision makers, actors, institutional stakeholders and power entities is of paramount importance to introduce successful changes. In many situations it may be necessary to create **ad-hoc structures that facilitate urban innovation**, for instance incubators, private-public partnerships, special purpose vehicles, ad-hoc institutions and many more. As we increasingly understand the relationship between smart cities and institutional arrangements, research will be needed to formulate models that can be replicated across cities and cultures, and provide a more sustainable basis for innovation adoption in cities.

From a quantitative viewpoint, the economic value of *Live City* services and applications can be either defined in **concrete revenues or as an after effect of improved quality of life**. The Economist Intelligence Unit's liveability ranking [40] quantifies the challenges that might be presented to an individual's lifestyle in 140 cities worldwide. Each city is assigned a score for over 30 qualitative and quantitative factors across five broad categories: Stability, Healthcare, Culture and Environment, Education, Infrastructure.

These five categories basically sum up 'what people want'. Interestingly, each of these categories can to some extent be improved or optimised by applying the principals of a *Live City* as described in this paper. In turn, improving a city's quality of life leads to a contented work force and families, and in turn, to increased economic value.

Most cites have not been planned from the ground-up and grew organically. The technologies that have been developed in the few last years, like pervasive sensors to assess urban dynamics and especially mobile technologies, offer new opportunities to 'tune' and 'fine-tune' the urban processes within cities, just as any other process can be optimised.

These urban processes can be transportation related, to monitor and direct the daily traffic in real time, optimise parking spaces and navigation to available parking, or simply to help people with their daily tasks, finding jobs, finding housing, connecting people in spare time, showing where less people are for leisure-time activities or where many people are for night life. Tools that bring the feedback loop directly to people make it easy to **promote events and give people instruments** to rate the attractiveness of these happenings.

Mobile technologies and the available development ecosystems offer great **opportunities for young start-ups** to build GPS-enabled, crowd-sourced, location-based apps. Just one example is the Wikitude World Browser [41], amongst 500.000 other apps, which are tailored at individual needs. Igniting and funding a start-up scene can be the starting point for any government to build a connected *Live City*: start-ups create jobs and apps, which in turn, if tailored for locals, benefit the people in the city and improve the quality of life.

The improved economic **value of a 'tuned' city**, i.e., better traffic management, optimised parking services, better housing and job search, city services and applications that deliver location-based news, events and happenings can be enormous. On one hand there can be cost saving advantages, for instance in considerable fuel savings if available parking spaces are reserved on a first-come-first-served policy and the driver is routed to this parking space rather than having to circle looking for a parking space.

A further important element in adopting smart technologies for urban innovation is the need to facilitate experimentation. The growth of the Internet industry in the last decade has brought about a model of companies largely based on small-scale experimentation, controlled failure and rapid adaptation. The idea that an internet company can be "planned and executed" (the waterfall model) has been replaced by the idea that an internet company "grows" through trial and errors while finding what works and what doesn't, accepting that **strategy is constructed ex-post**, rather than ex-ante. This modus operandi is being considered in many other sectors, which recognize the benefit of adaptability in the face of rapid external change.

Most cities, on the contrary, still work on long planning horizons and waterfall planning models are the norm. While cities cannot be compared to software projects, there are many areas of urban development that do not require infrastructure planning and investment and could be candidates for some radical review. The fact that we can increasingly measure the city in real-time and feedback the results of urban modifications in near real-time makes it conceivable to adopt trial-and-error policies on a much larger scale. While there are examples available, there is a strong need to develop **management frameworks** that would support organisations in this effort.

On the revenue side Google has shown in the last few years that Internet advertisement actually works. Google matches the search terms people enter in their search engine with ads. This works so well that it grew to 30 billion revenue per year, operating a million servers worldwide, serving 1 billion search requests every day. One key to

generating revenue in the field of *Live Cities* may be to apply what Google did with the Internet to the real world, offering information and search services that focus on time, location, context and people rather than on simply search terms.

VI. CONCLUSION

In opposition to projections, which stated that the wide-spread distribution of high-speed internet connections would render geographical distance irrelevant, cities have recently become the centre of interest in academic research. However, especially **real-time monitoring of urban processes** is widely unexplored and has recently received a lot of attention due to the fast rise of inexpensive pervasive sensor technologies, which made ubiquitous sensing feasible and enriches research on cities with uncharted up-to-date information layers.

Within this vision of a *Live City*, the city is not only regarded as a geographical area characterised by a dense accumulation of people or buildings, but more as a multi-layered construct containing multiple dimensions of social, technological and physical interconnections. Through this viewpoint of urban areas as an actuated **multi-dimensional conglomerates of dynamic processes**, the city itself can also be seen as a complex near real-time control system creating a feedback loop between the citizens, environmental monitoring systems, the city management and ubiquitous information services.

In the *Live City*, the everyday citizen is empowered to monitor the environment with sensor-enabled mobile devices. This feedback of 'sensed' or personally observed data, which is then analysed and provided to citizens as decision-supporting information, can change people's behaviour in how they use the city and perceive their environment by supporting their short-term decisions in near real time. However, this requires promotion of the user appropriation of the information through awareness of limitations.

Basically, we identified four main barriers towards the implementation of the *Live City* concept: methodological issues, technical/technological problems, lacking quantification of the economic benefits, and finally privacy and legislative questions. We discussed these challenges and highlighted **future research avenues** in Sections IV and V.

We believe that promoting the *Live City* concept will trigger a profound rethinking process in collaboration and cooperation efforts between different authorities. Also, a people-centric view of measuring, sharing, and discussing urban environments might increase agencies' and decision makers' **understanding of a community's claims** leading to proactive democracy in urban decision-making processes.

In terms of privacy and personal data collection, it is evident that everybody has to have the right to decide what kind of personal data is collected by whom, and for which purposes these data are used. In this context, people have to have an **opt-out possibility** to withdraw their consent to personal data collection. This is particularly important in the context of collective sensing, which tries not to exploit single people's measurements and data, but analyses aggregated anonymised data coming from collective networks, such as Twitter, Flickr or the mobile phone network.

Regarding the vision of digital earth, as formulated by Al Gore [42], both negative and positive aspects have to be addressed: positive aspects like emergency support, traffic congestion prevention, or, generally speaking, holistic real-time situational awareness of our environment; but also potential negative developments such as unwanted directed advertising, unauthorised tracking or extensive data mining have to be considered.

As mentioned in the Introduction, we are experiencing a fast progressing technology development, which is already moving ahead of society. The deciding final question can be: If we compare this development with a stream moving at high speed, on which we are paddling to remain on the same spot or at least not to drift off too fast, where does our goal for the future lie: down-stream, somewhere near our current spot, or even up-stream? In the end, legislation bodies are called upon to set the legal stage for leveraging *Live City* technologies, exploit economic opportunities, but still preserve citizens' privacy.

ACKNOWLEDGMENT

The authors would like to thank the audience of the panel discussion 'The Real-time City: Technology - Innovation - Society' at GI Forum and AGIT conferences (July 2011, Salzburg, Austria) for their valuable contributions to the discussion, which partly served as input for this publication. In particular, the authors would like to thank Prof. Dr. Manfred Ehlers and Prof. Dr. Petra Staufer-Steinnocher their fruitful comments.

REFERENCES

[1] Resch, B., Zipf, A., Breuss-Schneeweis, P., Beinat, E., and Boher, M. (2012) Live Cities and Urban Services - A Multi-dimensional Stress Field between Technology, Innovation and Society. In: Proceedings of the 4th International Conference on Advanced Geographic Information Systems, Applications, and Services - GEOProcessing 2012, Valencia, Spain, January 30 - February 4 2012, pp. 28-34.

[2] Cairncross, F. (1997) The Death of Distance: How the Communications Revolution Will Change Our Lives. Harvard Business School Press, Boston, MA, USA, 1997.

[3] Borsch, S. (2009) Distance is Dying. http://blogs.scholastic.com/accelerating_change/2009/07/distance-is-dying.html, 7 July 2009. (21 August 2011)

[4] Gilder, G. and Peters, T. (1995) City vs. Country: The Impact of Technology on Location. Forbes ASAP, 155(5), pp. 56-61, 27 February 1995.

[5] United Nations Population Fund (2007) State of World Population 2007: Unleashing the Potential of Urban Growth. United Nations Population Fund Report, ISBN 978-0897148078, New York, NY, UNFPA, 2007.

[6] Dierig, S., Lachmund, J., and Mendelsohn, A. (2000) Science and the City. http://vlp.mpiwg-berlin.mpg.de, Workshop, Max Planck Institute for the History of Science, Berlin, Germany, 1-3 December 2000. (10 September 2011)

[7] Netherlands Organization for Scientific Research (2007) Urban Sciences. http://www.urbansciences.eu, Interdisciplinary Research Programme on Urbanization & Urban culture in The Netherlands, 2007. (26 August 2011)

[8] SENSEable City Laboratory (2009) MIT SENSEable City Lab. http://senseable.mit.edu, September 2011. (14 September 2011)

[9] International Telecommunication Union (2010) Key Global Telecom Indicators for the World Telecommunication Service Sector. http://www.itu.int, 21 October 2010. (17 September 2011)

[10] Green, J. (2011) Digital Urban Renewal - Retro-fitting Existing Cities with Smart Solutions is the Urban Challenge of the 21st Century. http://www.cisco.com, Ovum Report OT00037-004, April 2011. (11 September 2011)

[11] European Commission (2012) European Initiative on Smart Cities - SETIS. http://setis.ec.europa.eu/about-setis/technology-roadmap/european-initiative-on-smart-cities, September 2012.

[12] ENoLL (2011) Open Living Labs | The First Step towards a new Innovation System. http://www.openlivinglabs.eu, September 2011. (11 September 2011)

[13] IBM (2009) A Vision of Smarter Cities - How Cities Can Lead the Way into a Prosperous and Sustainable Future. http://www.ibm.com, IBM Global Business Services Executive Report, 2009. (04 September 2011)

[14] Resch, B., Lippautz, M., and Mittlboeck, M. (2010) Pervasive Monitoring - A Standardised Sensor Web Approach for Intelligent Sensing Infrastructures. Sensors - Special Issue 'Intelligent Sensors 2010', 10(12), 2010, pp. 11440-11467.

[15] Murty, R., Mainland, G., Rose, I., Chowdhury, A., Gosain, A., Bers, J., and Welsh, M. (2008) CitySense: A Vision for an Urban-Scale Wireless Networking Testbed. Proceedings of the 2008 IEEE International Conference on Technologies for Homeland Security, Waltham, MA, May 2008.

[16] Townsend, A.M. (2000) Life in the Real-time City: Mobile Telephones and Urban Metabolism. Journal of Urban Technology. (7)2, pp.85-104, 2000.

[17] Foth, M. (Ed.) (2009) Handbook of Research on Urban Informatics: The Practice and Promise of the Real-Time City. ISBN 978-1-60566-152-0, Hershey, PA: Information Science Reference, IGI Global..

[18] General Services Administration (1996) Federal Standard 1037C. Telecommunications: Glossary of Telecommunication Terms, http://www.its.bldrdoc.gov, 7 August 1996. (11 September 2011)

[19] Resch, B., Mittlboeck, M., Lipson, S., Welsh, M., Bers, J., Britter, R., and Ratti, C. (2009) Urban Sensing Revisited – Common Scents: Towards Standardised Geo-sensor Networks for Public Health Monitoring in the City. In: Proceedings of the 11th International Conference on Computers in Urban Planning and Urban Management - CUPUM2009, Hong Kong, 16-18 June 2009.

[20] UCL Centre for Advanced Spatial Analysis (2011) Tweet-o-Meter - Giving You an Insight into Twitter Activity from Around the World!. http://www.casa.ucl.ac.uk/tom, 12 September 2011. (12 September 2011)

[21] Palmer, M. (2006) Data is the New Oil. http://ana.blogs.com, 3 November 2006. (12 September 2011)

[22] Kennedy, J. (2011) Data is the New Oil. http://www.siliconrepublic.com, 23 June 2011. (12 September 2011)

[23] Goodchild, M.F. (2007) Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. International Journal of Spatial Data Infrastructures Research, vol. 2, pp. 24-32, 2007.

[24] Craglia, M., Goodchild, M.F., Annoni, A., Camera, G., Gould, M., Kuhn, W., Mark, D., Masser, I., Maguire, D., Liang, S., and Parsons, E. (2008) Next-Generation Digital Earth: A Position Paper from the Vespucci Initiative for the Advancement of Geographic Information Science. International Journal of Spatial Data Infrastructures Research, vol. 3, pp. 146-167.

[25] Raper, J. (2011) Realising the Benefits of Open Geodata: Lessons from London's Experience. Keynote at AGIT 2011, 6 July 2011, Salzburg, Austria.

[26] Williams, M., Cornford, D., Bastin, L., and Pebesma, E. (2008) Uncertainty Markup Language (UncertML). OGC Discussion Paper 08-122r2, Version 0.6, 8 April 2009. (14 August 2011)

[27] Neis, P., Zielstra, D., and Zipf, A. (2012) The Street Network Evolution of Crowdsourced Maps – OpenStreetMap in Germany 2007-2011. In: Murgante, B., Borruso, G. And Gibin, M. (2012) Future Internet. Special Issue "NeoGeography and WikiPlanning", (DOI 10.3390/fi4010001), 2012(4), pp. 1-21.

[28] Helbich M., Amelunxen C., Neis P., and Zipf A. (2010) Investigations on Locational Accuracy of Volunteered Geographic Information Using OpenStreetMap Data. GIScience 2010 Workshop, Zurich, Switzerland, 14-17 September 2010.

[29] Roick, O., Loos, L., and Zipf, A. (2012) Visualizing Spatio-temporal Quality Metrics of Volunteered Geographic Information – A Case Study for OpenStreetMap. Geoinformatik 2012. Mobilität und Umwelt. Braunschweig. Germany.

[30] Hagenauer, J. and Helbich, M. (2012) Mining Urban Land Use Patterns from Volunteered Geographic Information Using Genetic Algorithms and Artificial Neural Networks. International Journal of Geographical Information Science (IJGIS), 26(6), pp. 963-982.

[31] Resch, B., Blaschke, T., and Mittlboeck, M. (2010) Live Geography - Interoperable Geo-Sensor Webs Facilitating the Vision of Digital Earth. International Journal on Advances in Networks and Services, 3(3&4), 2010, pp. 323-332.

[32] Botts, M., Percivall, G., Reed, C., and Davidson, J. (Eds.) (2007) OGC Sensor Web Enablement: Overview And High Level Architecture. http://www.opengeospatial.org, OpenGIS White Paper OGC 07-165, Version 3, 28 December 2007. (17 August 2011)

[33] Resch, B. (2012) On-the-fly Sensor Fusion for Real-time Data Integration. In: Löwner M.-O., Hillen, F. and Wohlfahrt, R. (2012) Geoinformatik 2012 – Mobilität und Umwelt, ISBN-10: 3844008888, Shaker Verlag, Braunschweig, Germany, pp. 229-236.

[34] Göbel, R. and Zipf, A.(2008) How to Define 3D Geoprocessing Operations for the OGC Web Processing Service (WPS)? Towards a Classification of 3D Operations. 12th International Conference on Computational Science and Its Applications (ICCSA), Perugia, Italy, 30 June-3 July 2008.

[35] Lanig S. and A. Zipf (2010) Proposal for a Web Processing Services (WPS) Application Profile for 3D Processing Analysis. 2nd International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2010), St. Maarten, Netherlands Antilles, 10-15 February 2010, pp. 117-122.

[36] Resch, B., Sagl, G., Blaschke, T., and Mittlboeck, M. (2010) Distributed Web-processing for Ubiquitous Information Services - OGC WPS Critically Revisited. In: Proceedings of the 6th International Conference on Geographic Information Science (GIScience2010), Zurich, Switzerland, 14-17 September 2010.

[37] Stollberg, B. and Zipf, A. (2009): Development of a WPS Process Chaining Tool and Application in a Disaster Management Use Case for Urban Areas. UDMS 2009. 27th Urban Data Management Symposium, Ljubljana , Slovenia.^

[38] Mayer, C. and Zipf, A. (2009) Integration and Visualization of Dynamic Sensor Data into 3D Spatial Data Infrastructures in a standardized way. GeoViz 2009 – Contribution of

Geovisualization to the concept of the Digital City. Workshop. Hamburg. Germany.

[39] Calabrese, F., Di Lorenzo, G., Liu, L., and Ratti, C. (in press) Estimating Origin-destination Flows Using Opportunistically Collected Mobile Phone Location Data from One Million Users in Boston Metropolitan Area. IEEE Pervasive Computing, 2011.

[40] Economist Intelligence Unit (2011) Liveability Ranking and Overview 2011. http://www.eiu.com, February 2011. (4 September 2011)

[41] Wikitude GmbH (2011) Wikitude World Browser | Wikitude. http://www.wikitude.com, September 2011. (13 September 2011)

[42] Gore, A. (2010) The Digital Earth: Understanding our planet in the 21st Century. Speech by Vice President Al Gore, Given at the California Science Center, Los Angeles, California, http://www.isde5.org, 31 January 2008. (4 September 2011)

A Generic Data Processing Framework for Heterogeneous Sensor-Actor-Networks

Matthias Vodel, René Bergelt, and Wolfram Hardt
Dept. of Computer Science
Chemnitz University of Technology
Chemnitz, GERMANY
Email: { vodel | berre | hardt }@cs.tu-chemnitz.de

***Abstract*—This paper presents a generic, energy-efficient concept for synchronised logging, processing and visualisation of arbitrary sensor data. The proposed framework enables a chronological coordination and correlation of information retrieved from different, distributed sensor networks as well as from any other self-sufficient measurement system. By relating data from different sensor data sources the overall information quality can be improved significantly. The implementation allows for easy cross-platform communication and facilitates readability by humans by employing the XML data format for data storage. Furthermore, the aggregated, heterogeneous sensor information can be converted into an extensible list of output formats. Depending on the application-specific requirements for visualization it is possible to consider additional meta-information from the test environment to optimise the data representation. The utilization of advanced data fusion techniques and pre-processing mechanisms enables a selective data filtering to reduce the network load. In order to evaluate basic usability requirements and the effectiveness of the proposed concept, an automotive sensor network is presented as a test system for the framework. For this demonstration, the available on-board measurement systems of a vehicle were extended by high-precision sensor nodes establishing a wireless sensor network and aggregating environmental and behavioural data on several test drives. Subsequently the correlated measurement information was converted and visualised for use with several professional data analysis tools, including jBEAM, FlexPro as well as Google Earth.**

Keywords-data aggregation; data fusion; data synchronisation; heterogeneous wireless sensor networks; sensor actuator systems

I. Introduction

Current research projects in the field of wireless sensor networks operate on different, proprietary hardware platforms and contain multifaceted types of sensors. The main objective of activities in this area is to establish an extensive knowledge base which concentrates data of different kinds and allows its user to gain further information out of the entirety of the collected data. In order to accomplish this goal, the fusion of data from different sources as well as methods for a uniform analysis are essential. This scenario is illustrated in *Figure* 1. Currently, most measurement scenarios consist of several application-specific and independently operating processes for the collection, storage and analysis of sensor data. There are no uniform synchronisation techniques between autonomous sensor systems in existence. Accordingly, a detailed and target-oriented post-processing of the data sets within a shared knowledge base is not feasible. In consequence, it requires much effort to create unambiguous, novel relations between different measurement information or this is simply not possible at all. Due to missing relations, it is very hard to create a common primary index for the given, heterogeneous sensor platforms.

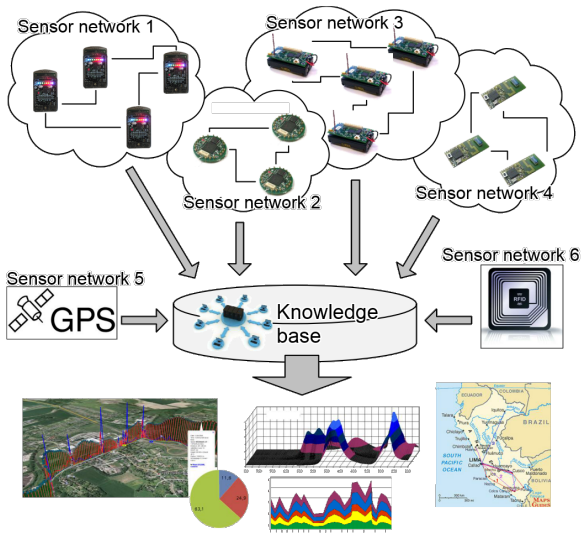


Figure 1: Gathering sensor data into a single knowledge base where it is analysed uniformly and independent of its origin

To solve this problem, we developed *GREASE* - a Generic Reconfigurable Framework for the Evaluation and Aggregation of heterogeneous Sensor Data [1], [2]. In order to introduce this integrated data processing concept, this paper is structured as follows: After this introduction, section *II* provides an overview about heterogeneous, distributed sensor environments, the data processing flow and respective challenges. The proposed GREASE framework is introduced in Section *III*, including conceptual fundamentals, basic requirements, system parameters and the top level structure (Section *IV*). Section *V* provides implementation details of the GREASE software architecture as well as the overall application flow within the framework. Section *VI* specifies

a sample application scenarios with all integrated components and environmental conditions. The corresponding data analysis is described and discussed in Section *VII*. Finally, the paper concludes with a summary and an outlook for future work in this research project.

II. Related Work

During the last two decades, a couple of commercial tools for measurement, data recording and monitoring were developed. Unfortunately, most of these tools have functional or conceptual restrictions. Some sensor hardware vendors offer exclusive, hardware-specific analysis tools, which require additional devices or do only cover specific product series. Other sensor systems do not offer any special software tools for extracting the measured data sets. Thus, there is no particular support for further post-processing steps.

In consequence, software solutions targeting this problem have been developed. These include *LabView* from National Instruments [3], *jBEAM* from AMS [4], or FlexPro [5], which offer multiple features to enhance the restricted vendor tools. These more general toolkits provide a larger set of generic data recording and handling functions. Furthermoere, the applications allow an interpretation of offline data from databases or files as well as the live analysis from a given data source. Both jBEAM and LabView operate platform-independently. The software applications support many established data formats and related communication interfaces. Especially jBEAM, which integrates the *ASAM* standard *(Association for Standardisation of Automation and Measuring Systems)* [6], enables an easy and modular extension with user-defined components whereas FlexPro already includes a lot of additional visual plugins and offers a complete visualisation framework for the given measurement data.

Nevertheless, the very high system requirements of all given software applications represent a critical disadvantage. Accordingly, these tools are not suitable for the usage in resource-limited data recording environments. Consequently, such frameworks have to be used in a second data processing step on dedicated workstations with sufficient hardware components and resources. Hence, small and energy saving hardware systems, which are used exclusively for collecting and storing multiple data from different sensor sources, are not able to use data aggregation and visualisation features of these software frameworks [7]. Due to these circumstances, most ongoing sensor system projects use proprietary software solutions to organise and synchronise the collected measurement data [8]. Additionally, in practice there are many critical compatibility problems between such software tools. In consequence, modifications of the measurement scenario or the system configuration take a lot of time and bind important resources. The user therefore demands a universal software tool for collecting and analysing the entire pool of sensor information in an application-specific and resource-efficient way. Automated or semi-automated data evaluation and visualisation techniques represent further essential requirements, especially for unattended long-term aggregations of environmental sensor data or test runs.

III. Concept

For the concept we had to find generic methodologies and standards to route information from different sensor systems into a common data processing unit in a synchronised way, considering scenario-specific configuration schemes and sensor parameters. In this respect, synchronised time stamps for the heterogeneous sensor data sets are very important in order to allow the correct identification of correlations in the common knowledge base later on. Furthermore, such techniques allow the definition of user-specific data analysis procedures during the measurement runtime. This also includes advanced data fusion techniques [9], [10], [11] to shrink the data volume directly within the sensor nodes or prior to storage.

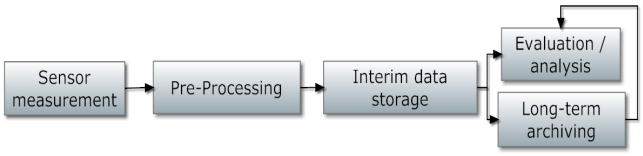


Figure 2: Data flow for typical measurement scenarios

The general data flow for (long-term) measurement scenarios is illustrated in *Figure* 2. Most available solutions only cover one half of this flow either the data aggregation or the evaluation of already collected and stored data. With GREASE we wanted to develop a framework which governs the whole data flow in such a measurement system in a generic and extensible way. To provide such features, GREASE represents a software framework based on a capable and lightweight data management concept, which is able to bypass the aforementioned mentioned disadvantages. It combines advanced sensor network configuration features with resource-efficient operating parameters. GREASE integrates the entire data processing flow, including all stages like data measurement, processing, storage and finally data analysis tasks. This flow is illustrated in *Figure* 3.

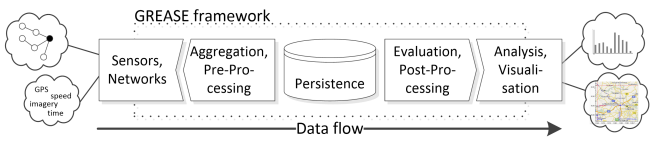


Figure 3: Integrated data processing flow for heterogeneous measurement topologies

Our concept focuses on resource-limited systems and has to be feasible for a wide variety of application scenarios. Thus, the primary objective is a dynamic and flexible

processing environment, which is adaptable to modifications in configuration or analysis requirements. Furthermore, the data processing core has to be separable into two spatial, chronological and platform-specific operating modes. All components for the data measurement are working within the first mode. All relevant modules for the data analysis as well as possible visualisation components operate independently within the second mode. Based on these dedicated modes, we are able to map different data processing functions to predefined configuration scenarios. In contrast, other related software tools do not separate the data handling process into such phases, especially with respect to the efficient concept of operations. Basically, GREASE acts as mediator between existing individual hardware sensors as well as sensor networks and evaluation processes including for instance visualisation tools. In this position it offers a standardised way to represent and transport measured data. The framework provides configurable synchronisation parameters for collecting information from several distributed sensor components. Accordingly, changes in the data analysis process do not affect the components of the data measurement and (in most cases) vice versa. This feature offers significant benefits, especially for complex sensor systems or inaccessible measurement environments.

In addition, all GUI *(Graphical User Interface)* actions, which can be accessed by the user, also have to be executable and controllable in an automated or semi-automated way. This feature represents another important difference to other related software tools, which do not provide any script-based operating modes without GUI. But especially for continuous maintenance-free and unattended sensor measurement scenarios, the scripting of user-defined activities is essential.

Hence, all central requirements for a synchronised data logging, processing and visualisation framework, particularly in the field of heterogeneous sensor network systems, can be summarised as follows:

- Synchronisation of different, autonomous sensor systems
- Modular extensibility through plugins and easy modification / adaptation
- Usage of a common data exchange format (e.g., *XML ((Extensible Markup Language)*) instead of proprietary data types
- Off-line and live data analysis from files, network file systems or databases
- Graphical User Interface for configuration and maintenance
- Automated or semi-automated data analysis and data representation mechanisms

Based on the proposed concept, the developed framework acts as coordinator between a given set of application-specific sensor and actuator components.

IV. Structure

As already mentioned, the structure of the proposed concept is basically divided into two operating phases. The first one encapsulates the data recording, synchronisation and correlation whereas the second phase provides processes for the data analysis generating a user-defined representation of the information sets. These two phases are connected through the so called persistence layer which is in charge of permanently storing the measured data in a well-defined format. This layer also restores data for future analysis and evaluation tasks. The described structure ensures a strict separation of data measurement and analysis tasks and is illustrated in *Figure* 4. Based on this approach GREASE exhibits a modular operating concept, making the whole framework independent of the given application-specific sensor configuration. Therefore, the environment uses an end-to-end communication design, called the *hourglass architecture*, which enables maximum interoperability between

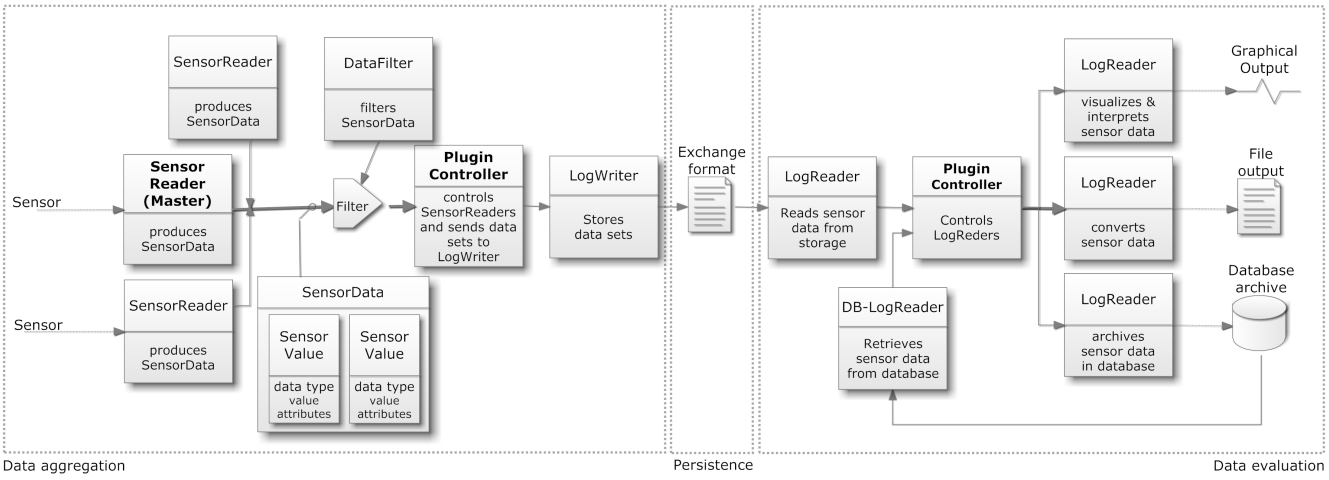


Figure 4: GREASE framework structure which allows the strict separation into two operation phases (measurement and evaluation)

the several components, i.e. modules or plugins. This results in a high diversity with respect to both components for data input (measurement) as well as data output (analysis). In contrast, the uniform connecting middle part is designed simple and light-weight. Correspondingly, the GREASE framework consists of data type definitions, interfaces and a plugin architecture. There exist five main types of plugins, three on the data input side (*SensorReader, DataFilter* and *SensorLogWriter*) and two on the data output side (*SensorLogReader* and *LogReader*). On both sides the coordination is handled by a dedicated *PluginController*. Every plugin type deals with exactly one of the phases illustrated in *Figure* 2 and *Figure* 3, respectively. Consequently, the SensorReader plugins are of particular importance as they transform measured sensor data of sensors into data types which adhere to the framework definition. They are able to combine several data values into data packets (*SensorData* objects) which consist of one or more *SensorValues*. A SensorValue includes at least sensor name, data type and the measured value and an arbitrary amount of descriptive meta-information may be added. These data packets are sent to the *controller* where *DataFilter plugins* enable the pre-processing of received sensor data. If any filter plugin has been registered at the *controller*, the packets will be forwarded to these filters in order to carry out further actions like editing or rejecting data. The *controller* then correlates the received data packets and will group them into datasets according to the scenario configuration. Within this process all internal and external parameters of the environment as well as special meta-information regarding the measurement scheme are correlated together with the datasets. Accordingly, researchers are able to reconstruct the whole test scenario with synchronised data, timestamps and a detailed system configuration. Therefore, the reuse factor, for instance in the field of automotive testing scenarios, increases substantially. In order to save the created datasets a single *SensorLogWriter* module has to be registered at the *controller*. Such a plugin allows the *controller* to write the datasets into a universal exchange format, e.g., an XML file. The gathered sensor data is now at the *persistence level* and may transferred to the evaluation part or archived for later assessment.

For the analysis of the sensor data, a corresponding counterpart of the used *SensorLogWriter* has to exist, a *SensorLogReader* which retransforms the data of the internal storage format into framework-specific data objects. This mechanism is - again - coordinated by a *PluginController*, which hands the retransformed sensor data to user-defined evaluation plugins (LogReader plugins). These plugins allow the user to evaluate, assess, visualise or archive the sensor data. A conversion to other file formats is also possible and can be used to import the measured data into third-party applications. An important advantage of the modular architecture is that the archived data, which has been saved for instance into a database, may be retrieved later on by an appropriate *SensorLogReader* plugin. Accordingly, the data can be processed by a second evaluation plugin. Due to the framework-defined data types, the evaluation modules do not know the origin of the sensor data (e.g., log file, database or any other storage type). Therefore, the development of a single plugin for GREASE can be carried out independently of other framework parts and future measurement scenarios. The persistence level represents the link between aggregation and evaluation and is one basic requirement for the exchange and editing of arbitrary measurement configurations. Existing configurations may now be altered easily by simply adding or removing components (i.e. plugins) without affecting the data flow of the overall monitoring system. Thus, changes within the measurement components are possible without having to update data analysis components and vice versa.

In principle it is also possible to connect data measurement and analysis modules directly, without using the persistence level. This leads to a kind of real-time data processing scenario. It may be noted that GREASE has not been designed for real-time processing and in the current implementation real-time qualities or maximum delays cannot be guaranteed. Therefore, the use of the framework in real-time decision making systems or systems requiring similar qualities is neither advisable nor wanted at the moment, since this is not the main application focus of GREASE. The framework performs best in long-term data measurement projects and in the domain of resource-limited systems without user interaction during measurements.

V. Implementation & Application Flow

To enable a platform-independent operation, the proposed concept has been implemented in *Java*. Nevertheless, additional modules can be written in other programming languages depending on the used controller. A precondition for both high flexibility and easy extensibility is a common interface specification within the controller. This feature of the proposed framework implementation is provided in the form of a *central core library*, which encapsulates the entire data processing logic. The communication between the controller and its set of modules is realized through dedicated protocols and a common syntax, which are both designed as generic as possible to allow a universal and flexible usage. This flexibility also simplifies the integration of third party modules and ensures the compatibility during further developments [12]. Due to the fact, that the sensor configurations and all kinds of scenario parameters are also transmitted within the XML representation, possible enhancements for customer-specific applications can be realized in an easy way.

Regarding the application flow, the initialisation of the *controller* commences with loading an application-specific

Table I: Supported data types by GREASE version 1.3

Type	Description
INT	Simple integer
FLOAT	Decimal
DATE	A date value without time information
TIMESTAMP	A timestamp (date & time information)
STRING	Arbitrary characters
IMAGE	Image data

configuration file. This file contains all information for the current project as well as the structure of all corresponding *SensorReader* components. Accordingly, the *controller* loads and activates all necessary sensor components and starts the data recording. Each *SensorReader* module operates simultaneously as a dedicated thread. When a *SensorReader* receives sensor data, a *SensorData* object will be generated. The format of this object is predefined by the framework configuration scheme. Such a *SensorData* object consists of a descriptive name (tag), e.g., *gps* for location data, and a list of *SensorValue* objects. A *SensorValue* object describes a single measured value and includes a descriptive name, the data type of the value, the data value itself and attributes which may provide further information about the value, for instance the physical measurement unit or special indicators. The latest GREASE version (GREASE.Core version 1.3) supports the data types listed in *Table* I.

The created *SensorData* object is sent to the *controller*, which forwards it to the registered filter plugins (only if available). Subsequently, it correlates and groups the received *SensorData* objects into datasets. In the following step, the object will be permanently stored through the persistence plugin the user has chosen. Due to the fact, that the *controller* has no knowledge about the interpretation of the transported sensor data, this process is fully application-independent. In contrast, jBEAM for instance uses project files which have to be adapted to changes in the measurement scenario since data measurement and analysis directly build on each other.

The standard storage plugin of GREASE saves datasets in an XML representation. The XML format has been chosen mainly for test and demonstration purposes as it also allows for human-readability. However, the user may replace it by any other third-party or custom plugin which uses different file formats or storage methods altogether. In more resource limited environments and applications this is advisable, since XML exhibits a rather high verbosity and leads to comparatively large file sizes.

Furthermore, the whole framework supports the translation and localisation of content and provides corresponding interfaces to plugins. In the current version, all standard plugins support both the English and German language. The basic setup of GREASE, including all plugins, is listed in *Table* II.

Table II: All plugins which are part of the basic setup of GREASE

SensorReader plugins	
MTS310Reader	Retrieving sensor data from Crossbow MTS310 sensor boards
GpsReader	Retrieving location data through the NMEA 0183 standard
ObdReader	Querying data from the OBD (on board diagnosis) interface of vehicles
CamImageReader	Capturing the image stream of image devices such as webcams
UdpSensorReader	Retrieving sensor data over UDP
DataFilter plugins	
LiveDataViewer	Visualisation of the data flow of the framework
NetVis	Visualisation of sensor network activity
Persistence plugins	
XML-LogWriter	Saves datasets as XML representation
XML-LogReader	Reads datasets from XML representation
DB-LogReader	Reads datasets from a database
LogReader plugins	
LogFuse	Fusing multiple log files
CSVOutput	Converting datasets to CSV format
ImageExtractor	Extraction of image data
DBOutput	Archiving of log files to a database
GEarthOutput	Visualisation of sensor data in Google Earth

With respect to the communication tasks in the sensor environment, we also have to discuss security features. Due to the fact, that GREASE focuses on research and development environments, we actually do not consider further security aspects for the distributed handling and storage of the sensor data. Within the different development stages of a given system, engineers design and implement complex test environments for getting valid and high-quality measurement results. Accordingly, the risks, which result from general communication threats are negligible. Nevertheless, we currently cooperate with related German car manufacturers in regards to this weak point. Several research projects focus on the development of advanced, energy-efficient security features for embedded, resource-limited sensor network topologies. In this context, the main challenge is the maintenance of a lightweight software architecture, which provides stable and flexible modules for diversified application scenarios. Here, advanced security mechanisms have a direct impact on the data throughput and the resource consumption. Accordingly, our goal is to find a good trade-off between runtime performance and security capabilities within GREASE. A first approach could then be the usage of an encrypting persistence module in non-time-critical applications. However, this leads to other questions as to which encryption algorithms to use and how to handle key security. The next step would then be to actually secure the data traffic inside the GREASE framework and its surrounding infrastructure.

VI. Application Scenario

The proposed concept has been developed to manage several sensor network scenarios at our computer engineering

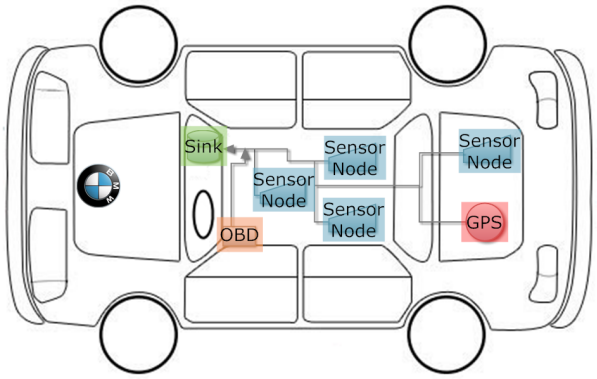


Figure 5: Measurement system - All data from the sensor nodes and the GPS module is transmitted to the data sink in the vehicle

```
<?xml version="1.0" encoding="UTF-8" ?>
<sensorlog targetProjectID="12" started="Fri May 25 12:37:12 CEST 2012">
    <dataset id="0">
        <gps friendlyName="GPS-Sensor" synchronized="true">
            <longitude type="FLOAT">12.9274235</longitude>
            <latitude type="FLOAT">50.839626</latitude>
            <speed type="FLOAT" unit="kmh">0.0</speed>
            <timestamp type="TIMESTAMP">Fri May 25 12:37:13 CEST 2012</timestamp>
        </gps>
        <obd friendlyName="OBD-SensorReader">
            <CalculatedEngineLoadValue type="FLOAT" Unit="%">59.21</CalculatedEngineLoadValue>
            <EngineCoolantTemperature type="FLOAT" Unit="°C">82.0</EngineCoolantTemperature>
            <EngineRPM type="FLOAT" Unit="rpm">1704.0</EngineRPM>
            <VehicleSpeed type="FLOAT" Unit="km/h">35.0</VehicleSpeed>
            <IntakeAirTemperature type="FLOAT" Unit="°C">36.0</IntakeAirTemperature>
            <ThrottlePosition type="FLOAT" Unit="%">25.88</ThrottlePosition>
            <RuntimeSinceEngineStart type="FLOAT" Unit="s">271.0</RuntimeSinceEngineStart>
            <FuelLevelInput type="FLOAT" Unit="%">8.62</FuelLevelInput>
        </obd>
        <cam friendlyName="CamImageReader">
            <img type="IMAGE" encoding="base64" quality="80"><![CDATA[<BILDDATEN>]]></img>
        </cam>
    </dataset>
    <dataset id="1">
    ...
    </dataset>
...
</sensorlog>
```

Figure 6: An excerpt from an XML log file produced by GREASE's standard persistence plugin

department. To clarify functional aspects of the implemented framework, we demonstrate the data processing flow using a real-world automotive measurement system [13]. For this monitoring scenario, the existing sensor components of a given research vehicle were upgraded with high-definition sensor nodes. These nodes are placed at predefined positions to monitor the entire environment and provide independent measurement data about the current temperature, light intensity as well as the acceleration in two axes and the magnetic field strength. Thus, the established wireless sensor network provides information about the measurement environment and external parameters. *Figure* 5 illustrates this measurement scenario.

The wireless sensor communication infrastructure is based on the *IEEE 802.15.4* and *ZigBee* [14][15] standards. Additionally, mobile sensor nodes are worn by the passengers. In order to realise localisation features for these nodes, they are equipped with *nanoPAN* ultra-low power network interfaces [16], which provide *RSSI*-based *(Received Signal Strength Indication)* distance information. Both communication technologies use the 2.4 GHz frequency spectrum for data transmission. A multi-interface, multi-standard data sink is able to handle both communication standards simultaneously. Robust communication stacks with adapted layer 2 and layer 3 protocols minimise interference-based influences on the communication behaviour.

In addition, we integrated a high-resolution *GPS (Global Positioning System)* sensor, which enables the correlation between absolute positioning information, speed, altitude and the available on-board vehicle data. We also established a connection with the vehicle's *OBD (On-board diagnostics)* interface in order to retrieve further information specific to the manufacturer and vehicle model.

By providing a synchronised knowledge base of all sensor information, a detailed analysis of specific driving situations and the driver behaviour is possible. Thereby, the GPS data allows a verification of these situations based on available track information. Accordingly, we are able to calculate and predict driver profiles. The results are used to adjust and to optimise the characteristics of the entire vehicle, for instance the engine management system or the suspension dynamics. This leads to the creation of in-depth driver profiles to adjust said car characteristics even prior to the engine's ignition. For this a driver must already be known to the system and enough data needs to be collected about their driving style beforehand. By fusing the location data with other sensor readings it is also possible to link peculiar measurement values with similar driving situations. This data can then be used to detect or predict similar route characteristics based on information retrieved from the knowledge base and thus critical situations may be identified more easily when assessing real-time data and appropriate countermeasures can be taken more quickly. Furthermore, an analysis of the wear measuring quantity provides interesting statements about the vehicle lifetime. For instance alterations in both the noise and vibration behaviour of components in the engine bay of a vehicle may indicate a damaged or soon-to-break part when compared to information in the knowledge base.

For the measurement environment described here, particular *SensorReader* modules for the data sink communication interfaces were implemented. Incoming data from the sensor network is classified and converted into abstract data objects, which are transmitted to the *controller*. Another *SensorReader* module implements the necessary features for the GPS (Global Positioning System) data input. Thereby, the *NMEA 0183 (National Marine Electronics Association 0183)* protocol for the positioning data is put to use. Accordingly, all kinds of GPS hardware, which support the NMEA protocol and the serial port as communication interface, are supported.

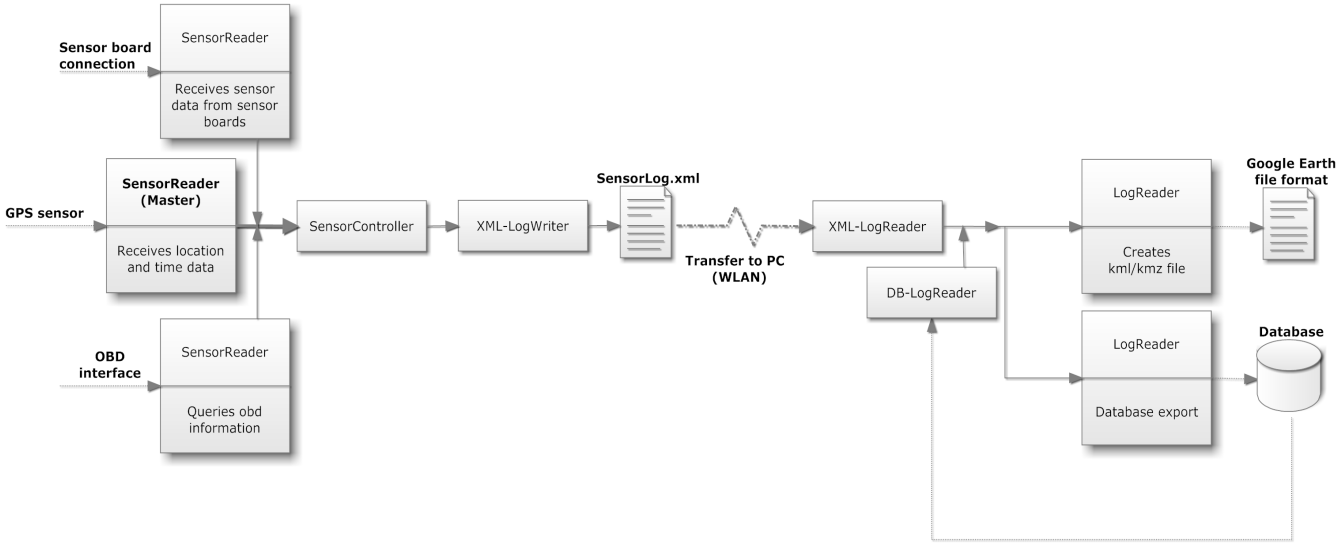


Figure 7: Sample application scenario of the proposed framework.

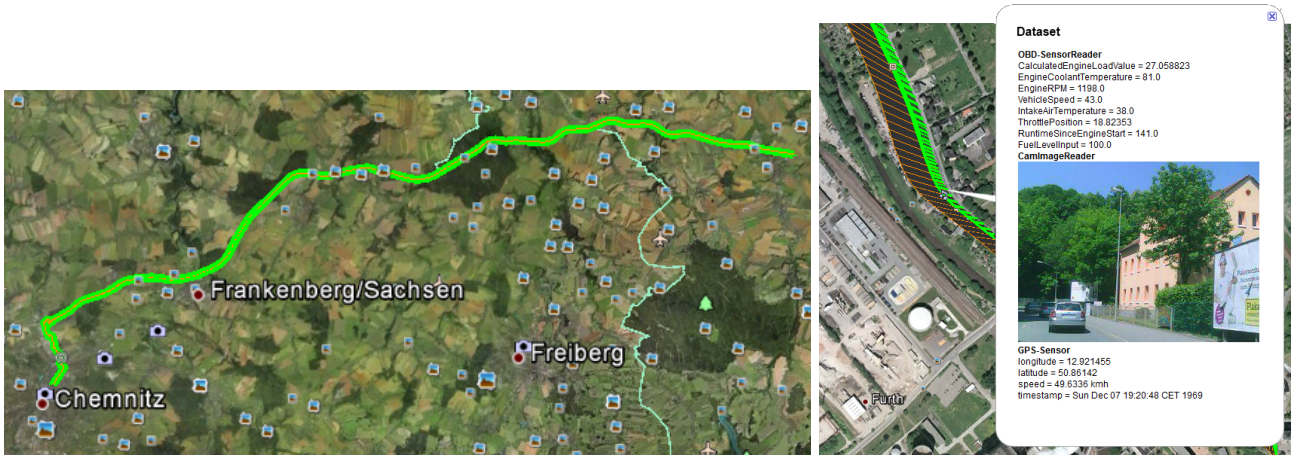
For the synchronisation of the sensor data, one specific *SensorReader* has to be predefined during the initialisation of the measurement scenario. In our example case, the system contains a GPS unit which provides positioning information as well as an accurate time signal. In consequence, the given timestamps from the GPS sensor represent the global *synchronisation master*. However, the framework is not restricted to using a timestamp as master index. Especially for the integration of multiple, autonomous sensor systems and a missing central scheduling entity, a user-defined choice for the synchronisation master provides important benefits. For instance this could be the activity of a specific sensor which detects predefined events or starting selective aggregation when certain sensor values exceed predetermined thresholds. On the one hand this helps in reducing the amount of sensor data to collect and store, whereas on the other hand it allows to group pieces of information which correlate based on time or event-wise. The storage of the retrieved datasets is handled by the standard persistence plugin of GREASE, which produces an XML file. An example of such a log file can be seen in *Figure* 6.

For post-processing and evaluating the collected data, two data analysis components were implemented. The first one is an export module, which prepares the sensor data sets for the storage in a given database system and accordingly transmits the chosen information. A second module is responsible for converting the sensor data with dedicated visualisation plugins, e.g., for Google Earth. Hence, the data output of this module is a *KML* (Keyhole Markup Language, file format of the Google Earth software) representation of all correlated sensor information or KMZ (zipped KML including resources such as images), if additional data has to be included. *Figure* 7 describes the data flow in this specific scenario. All developed modules are as generic as possible and may be used for a broad range of possible application scenarios and hardware-configurations in the future. This also includes a high compatibility level for both hardware and software environments [17][2]. For a detailed evaluation the import of measured data into JBEAM is possible and reasonable. This can be accomplished by using either the *CSV* (*Comma Separated Values*) export module or a specific jBEAM plugin, which implements the ASAM standard. With the latter the log file data can be imported directly into jBEAM which makes it even easier to assess and visualise large amounts of measured data in a comfortable and appealing way.

VII. Data Analysis

The following figures show some visualisation possibilities which represent the basis for further analyses. However, the actual evaluation and interpretation of the sensor data collected in the scenario described in *Section* VI is not part of this paper. Therefore, the figures may only be seen as examples with respect to the wide range of evaluation and visualisation possibilities the framework provides.

The Google Earth export module offers the conversion of the sensor data into the Google Earth file format (KML) which allows the user to use the Google Earth application for further analysis of the data. For this process, every dataset has to contain positioning information which can be retrieved from a respective sensor (for instance a dedicated GPS module). The module then creates a corresponding route for the GPS data (see *Figure* 8a) and a waypoint for every dataset. If the distance of two waypoints is smaller than a given threshold they may be fused for the sake of clarity. The user is now able to get detailed information for all waypoints in a corresponding pop-up window (see *Figure* 8b), this includes all sensor values contained in this



(a) The whole route (b) The details of a waypoint

Figure 8: Visualisation of a driven route and a single waypoint in Google Earth

dataset and any attached meta-information such as physical units or environmental circumstances. Furthermore, sensor value curves, such as the course of the vehicle speed, can be shown alongside the route as an additional three-dimensional altitude track. The altitude values of such a curve represent the respective sensor values as shown in *Figure* 9. Based on the timestamp data in the log file, Google Earth allows the user to view the course of the measurement by using its time bar. Only the appropriate waypoints for the selected time range will be shown. By animating this process the user can be provided with a comprehensive impression of the test drive. Additionally, each measurement scenario can be separated into different parts, for instance based on times when the vehicle did not move for a specific amount of time or if two adjacent waypoints have a distance greater than a defined value.

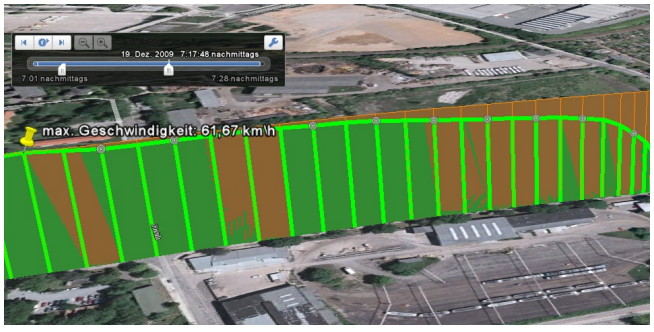


Figure 9: Value curves along the driven route in Google Earth

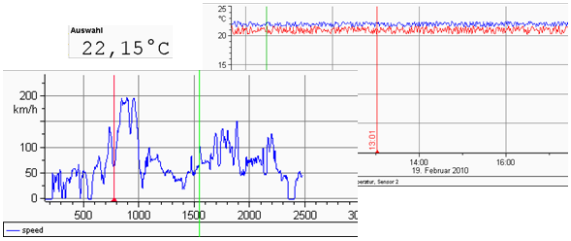
Other commercial software tools, e.g., FlexPro or jBEAM, are able to import the sensor data either by using export plugins such as the default CSV export module or through application-specific plugins which can the storage format produced by a GREASE persistence plugin. These third-party tools allow advanced, sectoral data post-processing tasks and offer advanced possibilities for the statistical analysis as well as the visualisation of the collected sensor data. Yet, the GREASE data flow and data measurement tasks remain unaffected regardless of the software tool, which shall be used for further tasks. A further sensor visualisation within jBEAM is shown in *Figure* 10a. In this figure there are two graphs which contain speed versus time and temperature versus time curves of a test drive. With jBEAM, classical value curves as well as maps with overlay information (as shown in *Figure* 10b) can be generated very easily. The entire data processing flow integrates all proposed features for the data recording, handling and visual representation. Since GREASE's aggregation phase and data flow are completely independent of the successive data evaluation the user can still benefit from the great variety of already existing software tools for this task. They may even change the used solution after already being halfway through a measurement project without problems or use several of such tools alongside each other. Besides, the framework also allows for a great flexibility for instance in adopting ever-changing measurement requirements and exposes a great reusability for upcoming projects.

VIII. Conclusion and Future Work

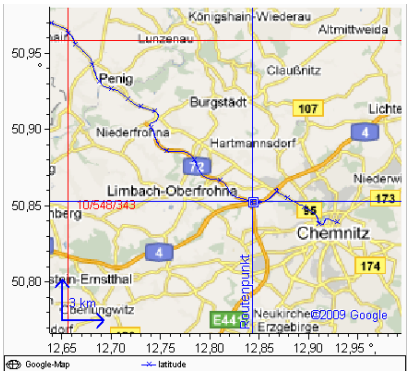
The proposed research work described the concept and implementation of a comprehensive data processing environment for heterogeneous sensor or sensor-actor-systems. The basic concept provides generic structures for many further research projects in the field of novel data aggregation and data fusion techniques. For an easy data collecting and data analysis process, we are now able to synchronise and correlate the single data sets also on resource limited and embedded computer systems. The result is a common and extensive knowledge base, which integrates all information sources into complex data sets.

In comparison to other related software tools, the proposed framework fulfils essential requirements for a flexible usage, a resource-efficient runtime behaviour as well as an automated or semi-automated operating mode. We developed a standardised process for monitoring and archiving data from a heterogeneous network topology in a synchronised way. Besides storing basic information from given sensor data sets, the system also integrates meta-information from the environment to increase the reuse factor of the measurement scenario. The universal XML data representation and a modular plugin system ensure a generic usage for all kind of sensor scenarios. Multiple data input and output interfaces provide a high level of compatibility to other software tools and data formats. The presented framework is used for several wireless sensor network projects at the Chemnitz University of Technology.

Regarding the presented automotive application scenario, the proposed framework enables correlations between the measured sensor data from the test track and specific driver profiles. Accordingly, these information allow dynamic adaptations of the driving parameters within the vehicle. This offers novel and interesting possibilities to optimise a vehicle for the specific characteristics of its driver.



(a) Value curves with markers for the currently selected waypoint



(b) A map with the driven route, showing the selected waypoint

Figure 10: Visualisation of sensor data in jBEAM

References

[1] M. Vodel, R. Bergelt, and W. Hardt. Grease framework - generic reconfigurable evaluation and aggregation of sensor data. In *Proceedings of the 2nd International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies (ENERGY2012 / InfoSys2012)*, page no pages given. IARIA, March 2012.

[2] M. Vodel, R. Bergelt, M. Glockner, and W. Hardt. Synchronised data logging, processing and visualisation in heterogeneous sensor networks. In *Proceedings of the International Conference on Data Engineering and Internet Technology*. Springer, March 2011.

[3] National Instruments. LabView. http://www.ni.com/labview/, 2010. [Online, retrieved: January, 2012].

[4] AMS GmbH. jBEAM. http://www.jbeam.de/german/produkte/jbeam.html, 2010. [Online, retrieved: January, 2012].

[5] Weisang. FlexPro. http://www.weisang.com/, 2010. [Online, retrieved: January, 2012].

[6] ASAM Consortium. Association for Standardisation of Automation and Measuring Systems. http://www.asam.net/, 2010. [Online, retrieved: January, 2012].

[7] L. Krishnamachari, D. Estrin, and S. Wicker. The impact of data aggregation in wireless sensor networks. In *Proceedings of the 22nd International Conference on Distributed Computing Systems Workshops*, pages 575–578. IEEE Computer Society, November 2002.

[8] M. Vodel, M. Lippmann, M. Caspar, and W. Hardt. Distributed high-level scheduling concept for synchronised, wireless sensor and actuator networks. *Journal of Communication and Computer*, 11(7):27–35, November 2010.

[9] V. Gupta and R. Pandey. Data Fusion and Topology Control in Wireless Sensor Networks. *WSEAS Trans. Sig. Proc.*, 4(4):150–172, 2008.

[10] H. Qi, S. S. Iyengar, and K. Chakrabarty. Distributed Sensor Fusion - A Review Of Recent Research. *Journal of the Franklin Institute*, 338(1):655–668, 2001.

[11] H. Qi, X. Wang, S. S. Iyengar, and K. Chakrabarty. Multisensor Data Fusion In Distributed Sensor Networks Using Mobile Agents. In *Proceedings of International Conference on Information Fusion*, pages 11–16, August 2001.

[12] A. Brown. *Component-Based Software Engineering*. Wiley-IEEE Computer Society Press, 1996.

[13] M. Vodel, M. Lippmann, M. Caspar, and W. Hardt. A Capable, High-Level Scheduling Concept for Application-Specific Wireless Sensor Networks. In *Proceedings of the World Engineering, Science and Technology Congress*, pages 914–919. IEEE Computer Society, June 2010.

[14] IEEE Computer Society. Part 15.4: Wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs). http://standards.ieee.org/getieee802/download/802.15.4-2006.pdf, 2007. [Online, retrieved: January, 2012].

[15] Zigbee Alliance. Zigbee specification. http://www.zigbee.org/en/spec_download/zigbee_downloads.asp, 2007. [Online, retrieved: January, 2012].

[16] Nanotron Technologies. Nanotron's transceiver enables iso compliant real time locating systems. In *The International Organization for Standardization and the International Electrotechnical Commission*, volume 24730-5:2010. New standard for Real Time Locating Systems (RTLS), April 2010.

[17] M. Vodel, W. Hardt, R. Bergelt, and M. Glockner. Modulares Framework fr die synchronisierte Erfassung, Verarbeitung und Aufbereitung heterogener Sensornetzdaten. In *Proceedings of the Dresdner Arbeitstagung Schaltungs- und Systementwurf*, pages 67–72. Fraunhofer Institute for Integrated Circuits, May 2010.

Dr. Matthias Vodel was born in Germany in 1982. He received the German Diploma degree (equal to M.Sc.) in Computer Science from the Chemnitz University of Technology with the focus on computer networks and distributed systems in 2006. In 2010, he received his Ph.D. degree in Computer Science from the Chemnitz University of Technology / Germany.
Currently, he works as a postdoctoral research fellow at the Department of Computer Science, Chair of Computer Engineering at Chemnitz University of Technology, Germany. His latest research projects focus on energy-efficient optimisation strategies in distributed embeddet systems. Additional fields of interest include network security topics, protocol engineering and wireless communication standards.
For his Ph.D. thesis, he received the "Commerzbank Award" for outstanding research work. He has published more than 40 journal and conference papers as well as two books. In 2008, Dr. Vodel received the best paper award for the conference paper "EBCR - A Routing Approach for Radio Standard Spanning Mobile Ad Hoc Networks". In 2012, the paper "WRTA - Wake-Up-Receiver Optimised Routing and Topology Optimisation Approach" received the best paper award at the 12th International Conference on ITS Telecommunications.
Dr. Vodel is member of the Association for Electrical, Electronic and Information Technologies (VDI/VDE) as well as IEEE member.

Dipl.-Inf. René Bergelt was born in Germany in 1988. He received the German Diploma degree (equal to M.Sc.) in Applied Computer Science with focus on embedded systems at the Chemnitz University of Technology in 2012. Currently, he is a Ph.D. student at the Department of Computer Science, Chair of Computer Engineering. The main fields of his research are methods for energy-efficient data aggregation and data fusion in wireless sensor networks. He also works on automotive applications for wireless sensor networks in the area of vehicle-to-vehicle and vehicle-to-environment communication strategies.

Prof. Dr. Wolfram Hardt is professor for computer science and head of the Computer Engineering Group at the Chemnitz University of Technology. He was born Germany 1965 and received the German Diploma degree (equal to M.Sc.) in Computer Science in 1991 from the University of Paderborn. Accordingly, Prof. Hardt received the Ph.D. degree in Computer Science from the University of Paderborn in 1996.
From 2000 to 2002 he was chair of the Computer Science and Process Laboratory at the University of Paderborn / Germany. Since 2003 Prof. Hardt became chair of the computer engineering Dept. at the Chemnitz University of Technology / Germany. He is editor of a scientific book series about self-organising embedded systems and has published more than 100 papers.
Prof. Hardt is member of the Association for Electrical, Electronic and Information Technologies (VDI/VDE), the Association for Computer Science (GI) and the Association for Computing Machinery. Since 2006 he is committee member of the DATE conference - "Design Automation and Test in Europe". His research interests include Hardware/Software Co-Design and Reconfigurable Hardware.

Beyond the Zermelo-Fraenkel Axiomatic System: BSDT Primary Language and its Perspective Applications

Petro Gopych
Universal Power Systems USA-Ukraine LLC
Kharkiv, Ukraine
pmgopych@gmail.com, pmg@kharkov.com

***Abstract*—A formalization of the recently proposed infinity hypothesis implying the common coevolution of the universe, life, mind, language, and society gives a possibility to introduce strict definitions of meaning and subjectivity, which spread beyond the traditional mathematics. This hypothesis leads to a semantic mathematics that is an implementation of the von Neumann's idea of a low-level "primary language" (PL). In this paper the formalization of this infinity hypothesis is further developed and some of its consequences are considered. In particular, a phenomenology formalization, definite and conditional meanings of the PL's words, their meaning complexity, categories and subcategories (hierarchies) of meaningful words, a way of the presentation of real numbers, the Cantor's continuum hypothesis, the PL's continuity-discreteness unity and uncertainty, non-Gödelian arithmetization by natural numbers and its relation to Chaitin's Omega-numbers, convention on truth, meaning ambiguity of words of different meaning complexity and its relation to Burali-Forti paradox are discussed. A validation of the PL is given. Some examples of meaningful computations using the technique of recently developed binary signal detection theory (BSDT) and the BSDT PL's perspective applications to solving the problems concerning the brain, mind and their faculties are briefly considered. It is emphasized super-Turing computations are typical for the BSDT PL as well as animal and human regular everyday meaningful communication.**

Keywords-infinity; meaning; subjectivity; phenomenology; context; categorization; attention; randomness; complexity; continuity-discreteness unity and uncertainty; arithmetization; continuum hypothesis; super-Turing computations.

I. INTRODUCTION

Recently proposed new infinity hypothesis [1] provides a possibility to strictly define such basic properties of mind as meaning and subjectivity. This hypothesis, contrary to the belief of some mathematicians [2], favors the view of mathematics as an invention of the mind. But mathematics is not only a product but also an instrument of the mind needed by humans to symbolically describe the world to better adapt to it. Mathematics may only be required and may only become possible in socially developed groups whose members are able to cooperate by means of a rather complex symbolic communication system or, in other words, by a language. In fact mathematics is an intrinsic part of the natural language (its fraction of maximal certainty) and can not be considered as something unrelated to it. Hence, mathematics as well as language is eventually the product of a particular human society and its culture, e.g., [3].

While humans are directing their efforts to mathematical problems which are in essence *external* with respect to their minds and faculties of minds (language, intuition, creativity, sociality, etc.), mathematics may be conceived, developed, and successfully applied in a completely *formal* way, ignoring the fact that it is inseparable from the mind/ meaning/subjectivity – it has been the course of the development of mathematics during thousands of years of its history. Rather recently this history was culminated in the design of formal axiomatic systems for mathematics as a whole – a finite number of most basic statements or *axioms* from which all mathematical theorems (correct assertions) can be derived in a finite number of logical inferences. This approach is known as a "finitist" one. The most famous and practically important of axiomatic systems is the Zermelo-Fraenkel (ZF) axiomatic system with the axiom of Choice (ZFC) [4]. The ZFC is widely recognized as a "standard" or "traditional" basis for all the contemporary mathematical formalism – a technique of writing out the axioms, theorems, and inference rules in a symbolic way. The very idea of the axiomatization and the goal of the famous David Hilbert's program [5] are to exclude from mathematics even the smallest traces of the mind/subjectivity, to reduce in this way mind-related ambiguities and to ensure as a result the highest possible (in the ideal case "absolute") logical rigor of it. But, as was demonstrated by Kurt Gödel [6], this enormous goal can never be achieved: in fact no finitist axiomatic system exists that leads to mathematical formalism that would simultaneously be consistent (all of its theorems do not contain logical contradictions) and complete (all of its theorems can be proved or, in other words, formally derived from the axioms). For the Hilbert's program and the whole axiomatic approach, this Gödel's incompleteness plays a destructive role. In spite of that, all theorems already proved and those that would only be proved in the future within the framework of the ZF/ZFC formalism remain valid while we are interested in problems that are not directed to our minds. An overwhelming majority of mathematical problems were till now exactly of this kind and, consequently, Gödel's incompleteness exerts no effect on them. Computational abilities of mathematical formalism that is based on ZFC-like axioms and ignores the mind were revealed and implemented by Alan Turing [7] as his famous abstract Turing machines.

The situation changes drastically as soon as humans start to address problems which are in essence *internal* with respect to their minds and faculties of minds. Now the minds have to symbolically describe themselves by means of methods created by them in a way that is comprehensible to other minds. Under such circumstances of self-reference, self-representation and self-awareness, standard (ignoring the mind) mathematics does not work and the impossibility of describing the mind by methods ignoring the mind manifests itself in some notorious paradoxes [4] and in Gödel's incompleteness [6] which becomes immediately practically significant. But even the most dogmatic formalists can not completely exclude the mind from their theories because the need always remains to explain (interpret) their special symbols in words of a natural language that names our elementary subjective experiences ("primitive intuitive knowledge"). As a result, mind-related meaning variability of natural languages penetrates into the mathematical formalism making it vague and insufficiently rigor. Henri Poincaré [8] formulated this problem as his "vicious circle argument" drawing the attention to the fact that at least some basic notions of known axiomatic systems are defined one through the other and, consequently, can not be treated as genuinely fundamental. Emphasizing the intuitionist aspects of knowledge, he also enunciated the intuitive origin of the fundamental in mathematics principle of induction. To solve these problems at least partially, formalists gather colloquial-word explanations of their symbols together, dub resulting collection the metamathematics, e.g., [9] and consider it as something different from the formalism itself.

There is also technically unpopular but methodologically important branch of mathematics known as L. E. J. Brouwer's intuitionism, e.g., [10]. It states that certain principal mathematical concepts (and, consequently, axioms as their symbolic representations) are immediately given to humans by their intuitions, though these concepts/axioms can never precisely be completed because over time they could be changed by further intuitions. Hence, the intuitionism and the formalism are axiomatic theories but with axioms introduced in different though similar ways (note, metamathematics, a part of the formalism, actually contains some elements of the intuitionism). It was John Lucas who first soundly stated [11] that the Gödel's incompleteness theorem [6] does not allow formal finitist explanation of the mind and presenting it as a Turing machine. Later these ideas were further developed by Roger Penrose [12], [13]. He, in order to explain mind/ consciousness, has argued the need of appealing to a so-far unknown hypothetical physics known as a "correct quantum gravity" that would be responsible for the emergence of our subjective experiences. In any case, Turing methods are insufficient to ensure mind/mind-related computations and to do this something "super-Turing" is required.

One can see, standard mathematics based on the ZFC or ZFC-like axiomatic approach becomes insufficient if we need to explain the mind and mind-related human/animal faculties. Numerous unsuccessful attempts to achieve an adequate description/understanding of such phenomena as mind/consciousness, e.g., [14], language, e.g., [15] or (mathematical) symbolism, e.g., [16] show these problems are tied in a Gordian knot and none of them, taken separately, could not fully be solved. It indicates we need a new mathematics spreading *beyond* the ZFC and equally successful in describing the phenomena that are *external and internal* with respect to the mind. Since standard mathematics successfully describes the mind-external world, it has to be a part of the required new mathematics. Hence, we need such a generalization of the ZFC that additionally takes into account a fundamental property of the world that is missed by the ZFC but crucially important for the emergence and maintenance of the human mind and mind-related faculties. We hypothesize [1], [17] this fundamental property is *the infinity of common "in the past" coevolution of the universe, life, mind, language, and society* (cf. Edward Wilson's idea of the "gene-culture coevolution" [18], Humberto Maturana and Francisco Varela's autopoiesis theory [19], Lynn Margulis's evolutionary symbiosis [20], and psychosomatic nets by Candace Pert [21]). In the present paper it is explained in which way this idea of infinity can be implemented by methods that are *beyond* the ZFC and regular Turing computations. The generality of our initial thesis entails the need to also address some other problems of great generality namely the context, meaning, attention, subjectivity, categorization, randomness, complexity, etc. In spite of that, this work is about science and not philosophy because it addresses a new mathematics and its practical computations.

For the sake of completeness, it is also needed to point to the opposite view usually accepted in the field of machine consciousness: the ZFC and Turing machines are sufficient for modeling the mind and serious obstacles that hinder achieving this goal are caused by severe but within the existing framework solvable technical problems, e.g., [22] - [24].

The rest of this paper is structured as follows. In Sections II to IV the hypothesis of concurrent infinity, based on it phenomenology formalization, and an implementation of the von Neumann's idea of a "primary language" (PL) are considered. It is also explained in which way the PL and recent binary signal detection theory (BSDT) [25] are closely related, why the BSDT PL spreads beyond the ZF/ZFC and why it may be treated as mathematics of meaningful computations. In Section V some details of the BSDT PL formalism are described, including the formalization of the notions of meaning, subjectivity and meaning complexity. Sections VI and VII describe a non-Gödelian arithmetization by natural numbers of all the BSDT PL expressions, their randomness and continuity-discreteness unity and uncertainty. The BSDT PL's convention on truth is considered in Section VIII. In Section IX the notion of conditional meaning is described and used to account for meaning ambiguity of BSDT PL words of different meaning complexities. A connection between the meaning ambiguity and Burali-Forti paradox is also demonstrated. Section X presents some numerical and empirical validations of the BSDT PL, including the existence in animals/humans of mirror neuron systems that implement typical super-Turing computations and BSDT PL super-Turing computers. In

Section XI examples of practically important given the context meaningful computations and some BSDT PL perspective applications are briefly discussed. Section XII gives conclusions.

II. BSDT PRIMARY LANGUAGE AS VON NEUMANN'S PRIMARY LANGUAGE

It was John von Neumann who was perhaps the first mathematician claiming the need of another mathematics for brain computations. This article is an attempt of an implementation of his idea of a low-level "primary language *truly* used by the central nervous system," and structurally "essentially different of those languages to which our common experience refer" [26, p. 92]. It is this primary language that is this new "essentially different" (we suppose, spreading beyond the ZFC) mathematics for describing the mind and doing mind/brain computations.

Since the PL is a low-level language for a nervous system's internal computations, its symbolism should be relevant to the usual style of signaling in nerve tissues of animals/humans by means of short electrical impulses of given amplitude often called "action potentials" or "spikes". We assume informative messages of interest are conveyed and processed in the brain as *patterns* of such spikes. It is supposed, these patterns are represented in the PL as finite-dimensional binary spin-like (with components ±1) vectors distorted by a non-additive "replacing" binary noise [27]. The coding by binary noise and, as a result, using the "one-memory-trace-per-one-network" learning paradigm [28] are the main features of the BSDT [25] that gives *the best* coding/decoding rules for patterns of binary signals damaged by binary noise. Complete description of all non-discrete properties of neurons as, e.g., their electric-chemical interactions or refractory periods is included into the infinite context giving the meaning to a particular pattern of spikes or respective binary vector (Section V B). They also contribute to uncertainties discussed in Sections VI B and X E. BSDT +1/–1 code can not be replaced by the 1/0 code traditionally used in most computers. It is superior because binary +1/–1, ternary +1/0/–1, and quaternary "colored" +1/–1 codes naturally describe 1) neuron assemblies in different states of their synchrony and 2) different ways of reciprocal ("phase") transitions between them [27].

To formalize the well-known vision of the brain as a *selectional* device, e.g., [29] within the framework of the BSDT, by analogy with Turing machines, we have introduced the abstract selectional machines, BSDT ASMs [30]. ASMs ensure *the best* BSDT decoding and give a technical implementation of the idea that *meaning* of a finite symbolic message is mainly defined by its *infinite context*. The BSDT and its ASMs became also the ground for the BSDT neural network assembly memory model, NNAMM [31], and BSDT atom of consciousness model, AOCM [32]. They employ explicitly the idea of the equivalence between the meaning of a message and *subjective experience* or *primary thought* of an organism (perceiving agent) recognizing this message. Such an approach inevitably requires an extending of standard mathematics beyond the ZFC, to ensure the consideration of the phenomenon of meaning/subjectivity/privacy at mathematical level of logical rigor. The latter is the mandatory prerequisite for solving what is called the "hard" problem of consciousness [33].

At the same time the PL would remain an empty enterprise if there is no technique implementing it computationally. Fortunately, the BSDT gives such *the best* technique that is completely ready to be used. What is additionally needed to ensure its success is a methodology of its application to particular PL-specific problems, see Section XI. Hence, for the PL, the BSDT plays a two-fold role: on the one hand, it contributes to its substantiation; on the other hand, it gives its computational implementation. Consequently, it is natural to refer to the PL we propose as the BSDT PL. The PL in turn gives the BSDT the significance of the best technique of the PL's meaningful computations (e.g., Sections IV, V, X, and XI).

In the regular sense of this term, proofs are understood as finite sequences of formal symbolic logical transformations that draw the theorems from axioms. For BSDT PL statements, such formal proofs have strictly speaking no sense because, in this case, we are always interested in their meanings but standard mathematical formalism rejects meanings by definition. For the substantiation of BSDT PL statements, we will give neither theorems nor proofs. We will provide instead their unambiguous *constructions*, given our new premises (Section V) and known theorems of standard mathematics. Such a style of writing is "constructive" rather than formal.

In order to arrive at the BSDT PL we have mainly been motivated by biological and mathematical reasons. For this reason, along all this paper we will focus on those problems of life, mind, language and society that standard mathematics fails to resolve. The most acute of them and most amenable for the first BSDT PL application are perhaps the reliable language communication *without syntax* in humans and communication *without any language* at all in human infants and animals of the same or relative species. These problems are of great importance for linguistics, cognitive sciences, and artificial intelligence because their study informs us about the dynamics of language as a population phenomenon, bodily forms of signaling, and about a cognitive and bodily infrastructure for social interaction [34]. We also highlight the ranges of BSDT PL applications and show where and in which way it could be reduced to the reining standard mathematics.

The BSDT PL is of course not restricted to biology; it may also be useful, e.g., in physics but this direction of BSDT PL perspective applications remains out of the scope of this work.

III. HYPOTHESIS OF CONCURRENT INFINITY, EBSDT AND BSDT PL PHENOMENOLOGY

The ZFC axiom of infinity postulates the infinity of the number of those elements/individuals that are used in ZFC theory of sets for the construction of these sets [4]. The meaning of the term "element/individual" is not specified in any way but it would be reasonable to believe (or at least one would prefer to believe) this infinity axiom reflects in a sense the tacitly assumed infinite richness of the world in which we

live in, though what is "the world" is again explicitly not specified. In spite of obscure terminology used we have to agree that the ZFC seems to imply the infinite versatility of the world but certainly does not inherently imply the possibility of its evolution and development. The ZFC world is a stationary one. This theory allows the allocation of different aggregates of elements (the world's "currently visible" fragments) but does not allow any changes of neither the world as a whole nor its currently visible parts. The fragments of ZFC world (sets and subsets of elements) are "tautologically" [4] related to each other like ZFC theorems/tautologies that could be transformed one into the other with the help of simple or intricate but always *reversible* (as they produce the *tautologies* only) formal rules. The irreversibility of known irreversible functions originates from a randomness of processes, e.g., [27] they represent but are not from the ZFC. If one prefers to keep the elements as abstract and not related to the world entities then ZFC mathematics remains meaningless. In spite of that its computations may gain different meanings from different, e.g., physical problems they describe (see Section XI B).

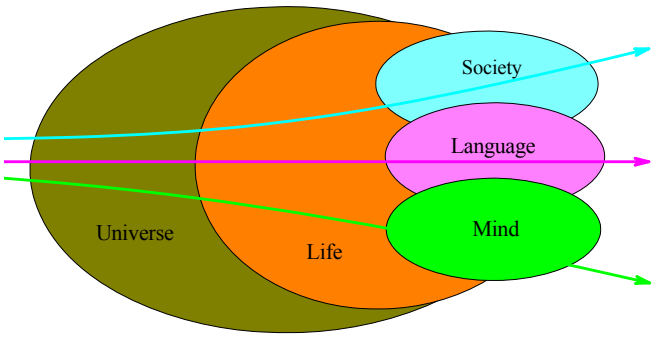


Figure 1. Hypothesis of concurrent infinity. Reciprocal relations between the universe, life, mind, language and society are shown as overlapping ovals of different colors. Arrows designate the course of their common infinite "in the past" (on the left) and open-ended "in the furure" (on the right) co-evolution.

The BSDT PL infinity hypothesis we are inaugurating is needed to introduce in mathematics the idea of permanent or "eternal" open-ended evolution and development of our (physical) world including animals/humans and their minds as a part of it. *In addition* to the ZFC-like infinite richness of the world, we postulate the infinity of common "in the past" and open-ended "in the future" co-evolution (Figure 1) of the universe, life, mind, language and society [1], [17], [32]. We also equate a *real-world physical device* devoted to the recognition of particular meaningful symbolic (binary for certainty) message originated from a thing of the world, complete binary *infinite on a semi-axis description* of the story of designing this device in the course of its infinitely long evolution from "the beginning of the world" until now, and *the meaning* of the message under consideration or *primary thought* it conveys. In addition *meanings* are interpreted as *subjective/first-person/private experiences* or respective *feelings* (*qualia*) and vice versa [1], [17], [32]. It is assumed, the BSDT PL world (it coincides with our physical world), is the total collection of what we call things, i.e., any inanimate objects, animate beings, and any relationships between/within them. All the world's things are permanently evolving, in the course of their common infinitely long *coevolution*, "in parallel" or *concurrently*, physically interacting with each other either directly or by their contributions to their common environment. To emphasize this issue we call our hypothesis *the hypothesis of concurrent infinity*. The BSDT extended by this hypothesis is referred to as *the eBSDT*; it is the basis for the BSDT PL we describe here as well as BSDT AOCM [32].

Our infinity hypothesis defines simultaneously a phenomenology (Figure 2), that is, explicit relationships between human subjective experiences and real world things humans perceive. The phenomenology is the branch of philosophy and science that emphasizes the role of/ concerns mainly with perceptual/subjective aspects of our knowledge and our minds. In its present form, it was established and strongly advocated by Edmund Husserl, e.g., [35]. For the recognition of the meaningful message of interest (a binary vector x_j^i given its infinite binary context c_{xi}, Section V), our phenomenology postulates the use of a real-world implementation of the BSDT ASM devoted to process the x_j^i ([30] and box 1 in Figure 2) and, consequently, it is *the BSDT PL phenomenology*. It not only connects our feelings to things we perceive (boxes 1 to 3, some details see in [32]) but ensures also their *formalization* giving a possibility (box 4) of establishing a set of formal rules defining in which way to deal with meanings and feelings in terms of standard mathematics (Section V). From a common-sense point of view, taken together, explicit phenomenological traits of our infinity hypothesis do give the theories based on them (BSDT PL and BSDT AOCM [32]) a little taste of "strangeness".

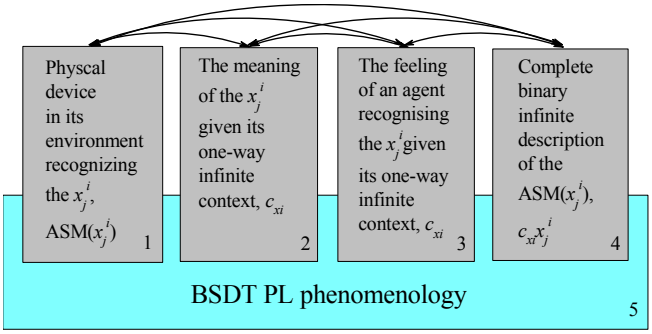


Figure 2. The BSDT PL phenomenology. Real-world physical devices recognizing the strings/vectors x_j^i and their equivalents in meaning, feeling, and symbolism domains (boxes 1 to 4) give rise together to BSDT PL phenomenology (box 5). Bidirectional arrows are used to designate the signs of equivalence and "paradigm shifts" or transitions "between incommensurables" [36, p. 150].

The distinctive feature of the hypothesis of concurrent infinity, which is rather difficult to comprehend, is that it *literary equates* usually *incommensurable* entities, related to quite different domains – real-world physical devices (box 1 in Figure 2), meanings of names of real-world things (box 2), subjective feelings (box 3), and infinite strings of symbols (box 4). In particular, our suggestion (boxes 1 and

2) that the meaning of a thing's name is identical to the physical device recognizing this name but not to the thing itself seems at the first glance counterintuitive. At least Ludwig Wittgenstein stated the opposite: "A name means an object. The object is its meaning." [37, 3.203]. I.e., he equated meanings of words and *the things* to be named while we equate meanings of words and *an agent's devices* recognizing these words. In other words, it is assumed the meaning of a thing's name is a specific internal activity or *psychological state* or *primary thought* of an organism (sensory agent) perceiving this thing. Name meaning is the property (momentary internal state) of a perceiving agent and not the property (feature, trait, state) of the thing to be named. It is easier to intuitively acquire this statement, if to remember that the things of the world are always given to us *indirectly*, through our sensory organs and respective patterns of sensory signals in our nerve tissues.

IV. The Idea of Phenpmenology Formalization and Semantic Mathematics

To do the formalization of BSDT PL phenomenology, we invoke our suggestion that an agent's psychological states, name meanings, primary thoughts and physical devices devoted to recognize the names are equivalent to infinite on a semi-axis binary strings that share their infinite on a semi-axis initial parts/beginnings or, in other words, that have common *prehistory*. It is a *pre*history and not the history because it describes the beginning of everything in the world and remains always essentially *unspecified*. We know about the prehistory that its existence is one-way infinite and common for all the things of the world but *nothing* more. Of this follows that a particular *infinite on a semi-axis* binary string describing the meaning of a thing's name (the physical device that recognizes this name and feeling the thing causes in a perceiving agent) has *infinite on a semi-axis* beginning that coincides bit-by-bit with *infinite on a semi-axis* beginnings of other *infinite on a semi-axis* binary strings describing the meanings of names of other things of the world. Consequently, to formalize the operations with name meanings and respective feelings (Section V), it suffices to fix such an arrangement of their one-way infinite meaning descriptions when their infinite on a semi-axis beginnings (initial parts or prehistory) coincide completely and, after excluding these common beginnings from the consideration, to deal with their *finite-in-length string remnants* only by methods of standard mathematics. Such mathematics of computations with *meaningful* one-way infinite and specially arranged strings we call *meaningful* or *semantic mathematics*.

The fact that meaningful/semantic computations with finite binary strings are defined given their common infinitely long on a semi-axis prehistory or under condition that their infinite on a semi-axis beginnings coincide bit-by-bit completely transforms them into a kind of *conditional computations*. Hence, BSDT PL computations obeying the demands of the hypothesis of concurrent infinity and respective BSDT PL phenomenology are ultimately the computations performed by methods of standard mathematics *given additional boundary conditions specified by a binary text that occupies completely an infinite semi-axis*. The BSDT PL as a semantic mathematics is *a generalization of standard mathematics* for the case of operations with one-way infinite binary strings having common one-way infinite beginnings and, simultaneously, a kind of *standard mathematics conditioned by* infinitely large amounts of additional assumptions written as an infinite on a semi-axis binary string. Once these additional boundary conditions (assumptions) are discarded, the mathematics of meaningful computations disappears and becomes the standard ZFC mathematics that by definition ignores meanings of its theorems/computations.

Since meaningful (semantic) computations are dealing with infinitely long strings/messages (i.e., taken as a whole genuine real numbers), they cannot be performed by regular Turing machines. To cope with semantic computations, a *super-Turing* computational technique and its implementation in the form of a physically constructible super-Turing computer [38] are obviously required (see Section X B and D). Like Turing machines implement the computations in standard ZFC or ZFC-like mathematics, super-Turing machines should implement the computations in semantic mathematics

V. Elements of the BSDT PL Formalism

The acceptance of the hypothesis of concurrent infinity transforms ZFC mathematics into the BSDT PL whose formalism differs to an extent from what we customary use.

A. Alphabet of Meaningful Words

As the BSDT PL is a kind of standard mathematics, the alphabet of the latter may be accepted as the alphabet of the former, with reservations concerning the specificity of semantic mathematics. The most important of them is that basic BSDT PL objects are *infinite* on a semi-axis symbolic (binary for certainty) meaningful strings that have common coinciding bit-by-bit *infinite* on a semi-axis meaningful beginnings. The other is that all these strings are written in the BSDT format as spin-like ±1-sequences and their finite-in-length end-fractions are processed by respective BSDT ASMs. In other words, it is assumed, all BSDT PL finite binary strings are coded using replacing binary noise [27] and decoded by BSDT ASMs [30] that give in this case the best decoding rules. The type of coding of infinite-in-length but explicitly unspecified common beginnings of meaningful strings does not matter. Meaningfulness of BSDT PL one-way infinite strings (and their equivalence to real-brain physical devices) is actually *postulated* by the hypothesis of concurrent infinity (Sections I and III, Figure 1) and respective BSDT PL phenomenology (Sections III and IV, Figure 2).

B. Vocabulary of Meaningful Words

The distinctive feature of the BSDT PL is that its basic elements (words) are meaningful. Their meanings are introduced as follows.

1) Meaningful simple words: Let us *arbitrary* choose one of BSDT PL one-way infinite binary strings, e.g., the c_{x0}

as a "master string". The length of it in bits, $l(c_{x0})$, equals by definition $\aleph_0$: $l(c_{x0}) = \aleph_0$ where $\aleph_0$ is the Georg Cantor's aleph-nought. If to divide the c_{x0} in the ith *randomly* chosen place into two parts then its finite and infinite fractions could respectively be thought of as an i-bits-in-length *simple meaningful word* x_j^i naming the ijth thing of the world and *the context* c_{xi} in which this word appears. Taken separately, x_j^i is meaningless. Resulting string $c_{x0} = c_{xi}x_j^i$ may be treated as the jth value of the string function/form $C(x^i) = c_{xi}x^i$ where string variable x^i is a string template of i empty cells needed to produce the strings x_j^i by filling this cell template in different i-length arrangements of +1s and −1s ($c_{xi}x^i$ is a concatenation of infinite binary string c_{xi} and cell template x^i). At a given value of i, with the help of the $C(x^i)$, 2^i of different strings $c_{xi}x_j^i$ can be generated with the same context c_{xi} and different affixes x_j^i. An affix x_j^i may simultaneously be treated as the ijth i-bits-in-length binary string, message, computer code/algorithm, vector or point in the space S_{xi} ($x_j^i \in S_{xi}$; $j = 1, 2, \ldots 2^i$), element of the set S_{xi} of the cardinality $|S_{xi}| = 2^i$, BSDT PL word or name (indices i and j point to a particular thing of the world, see Section V B3). Depending on the current context, these terms will further be used interchangeably.

By changing the values of i from zero to infinity, the function $C(x^i)$ allows to generate (construct) any finite binary string x_j^i of any length $i = l(x_j^i)$ given its infinitely long context c_{xi}. All resulting strings $c_{xi}x_j^i$ constitute together an *ultimate or proper class* S_{cx0} – the set that is not a member of any other set [39], $c_{xi}x_j^i \in S_{cx0}$. The term "proper class" may intuitively be interpreted "as an accumulation of objects which must always remain in a state of development" [40, p. 325]. The items of the S_{cx0}, $c_{xi}x_j^i$, are uniquely specified by their i-bits-in-length affixes (right-most end-fractions) x_j^i. The c_{xi} is common infinite context for all the x_j^i of the length i that have different arrangements of their ±1 components; $i = 0, 1, 2, \ldots$ and $j = 1, 2, \ldots, 2^i$. The principal property of elements of an S_{cx0} is that, in the sense of Cantor, they are all of the same infinite length $\aleph_0$ (are countable) but, in spite of that, they and their infinite fractions are explicitly *comparable* and may be a number of bits longer or shorter with respect to each other. Of the vantage of standard mathematics, the latter conclusion is fundamentally impossible though it is the norm for the BSDT PL due to the common infinite beginning of all its one-way infinite strings. For example, if meaningful simple words x_j^i and x_j^k obtained with the help of forms $C(x^i)$ and $C(x^k)$ given the same master string $c_{x0} \in S_{cx0}$ are of different lengths (e.g., $k > i$) then $l(c_{x0}) = l(c_{xi}x_j^i) = l(c_{xk}x_j^k) = l(c_{xi}) = l(c_{xk}) = \aleph_0$ but $l(c_{xi}x_j^i) - l(c_{xi}) = i$, $l(c_{xk}x_j^k) - l(c_{xk}) = k$, $l(c_{xi}x_j^i) - l(c_{xk}x_j^k) = 0$, and $l(c_{xi}) - l(c_{xk}) = k - i > 0$ (for $c_{xi}x_j^i$ and $c_{xk}x_j^k$, their the largest common infinite beginning is c_{xk}; see Figure 3).

Note, infinite words [41] of automata theory have no common infinite beginnings and remain within the framework of traditional mathematics

2) Meaningful composite words/sentences and their focal and fringe constituents: If string variable x^i consists of variables u^p and v^q then $x^i = u^pv^q$ with $i = p + q$; u^pv^q is a concatenation of cell templates u^p and v^q. The values of variables x^i, u^p, and v^q are respectively the strings x_j^i, u_r^p, and v_s^q that are the members of sets S_{xi}, S_{up}, and S_{vq} whose cardinalities are respectively $|S_{xi}| = 2^i$, $|S_{up}| = 2^p$, and $|S_{vq}| = 2^q$; $S_{up} \subseteq S_{xi}$ and $S_{vq} \subseteq S_{xi}$. The values of composite variable $x^i = u^pv^q$ are the composite words $x_j^i = u_r^pv_s^q$, the order of the composite word's constituents is essential for them (see an example in Figure 4). Composite variables consisting of any number of their constituents may similar be constructed. Composite space S_{xi} may also be interpreted as either the S_{up} whose vectors are colored in 2^q colors or the S_{vq} whose vectors are colored in 2^p colors. If so, then p and q are the measures of discrete "colored" non-localities of vectors in spaces S_{vq} and S_{up}, respectively [17], [27]. Similar colored (blue-and-red) binary spaces (three-dimensional "colored Boolean cubes") have earlier been used for representing the Boolean functions of one-dimensional cell automata, e.g., [42, ch. 6]. The rainbow of colors in finite-dimensional binary spaces discussed here is a direct generalization of two-color spaces of any dimensionality we previously introduced [27] to describe the coding of signals in nerve tissues of animals/humans.

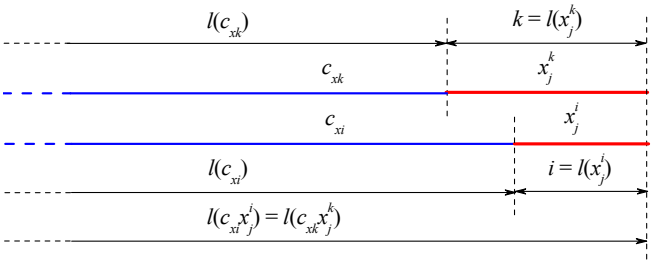


Figure 3. Meaningful simple words x_j^i and x_j^k (red line segments), their infinite on a semi-axis contexts c_{xi} and c_{xk} (blue line segments), and their meanings $c_{xi}x_j^i$ and $c_{xk}x_j^k$ (red and blue line segments taken together). The lengths of red line segment in bits, $i = l(x_j^i)$ and $k = l(x_j^k)$, are ensemble complexities of x_j^i and x_j^k ($k > i$), the lengths of blue line segments, $\aleph_0 = l(c_{xi}) = l(c_{xk})$, are their context complexities ($l(c_{xi}) - l(c_{xk}) = k - i > 0$); the lengths of red and blue line segments taken together, $\aleph_0 = l(c_{xi}x_j^i) = l(c_{xk}x_j^k)$, are their meaning complexities ($l(c_{xi}x_j^i) - l(c_{xk}x_j^k) = 0$, Section V C). Colored line segments denote the strings themselves, arrows designate their lengths. Dashed ending of lines on the left designate their infinity "in the past".

Isolated composite words are meaningless. Like simple words, they take their meanings from their infinite contexts and from themselves. For this reason, definite meanings have either whole composite words or their right-most fractions only. The right-most fraction of a composite meaningful word occupies an animal's dynamically created "focus of attention" and is called the "*focal*" word. The composite word's non-focal component is the focal word's "*fringe*" (by analogy with fringes of memory and consciousness [31], [32]) or the focal word's short-range or immediate or local context. If internal structure of a composite word is ignored then it is interpreted as a focal word that has zero-length fringe.

Meaningful composite words $x_j^i = u_r^pv_s^q$ are thought of as BSDT PL meaningful *sentences*. The focal word v_s^q corresponds to a sentence's feature/attribute that is currently in the focus of an animal's attention; its fringe u_r^p is the

fringe of the animal's memory or consciousness. A composite word's "holophrasical" presentation (without noticing its internal structure), e.g., x_j^i corresponds to the perception/understanding of a sentence as a whole whereas its "analytical" presentation as, e.g., a set of possible focal words v_s^q with $1 \le q \le i$ gives the sentence's meaning as a series of meanings of its simple focal words. Composite word's holophrasical presentation describes the perception of a thing as a whole (*diffuse* focus of attention) whereas its analytical presentation describes its perception as a series of its attributes (*acute* focus of attention). Any paraphrase of BSDT PL sentences (any other choice of their constituents) cannot change their whole meanings and in that sense the BSDT PL lacks "compositional semantics" [43]. Owing to our infinity hypothesis (Figure 1) and its phenomenology (Figure 2), meaningful BSDT PL sentences (meaningful composite words) are simultaneously real-brain devices processing these sentences. For this reason, the number of a composite word's constituents may be treated as an animal's *logical or reasoning deepness*. Since logical or reasoning deepness measured in humans is 3 to 5 [44], the biologically most plausible number of components constituting BSDT PL composite words is expected to be of the same value, 3 to 5.

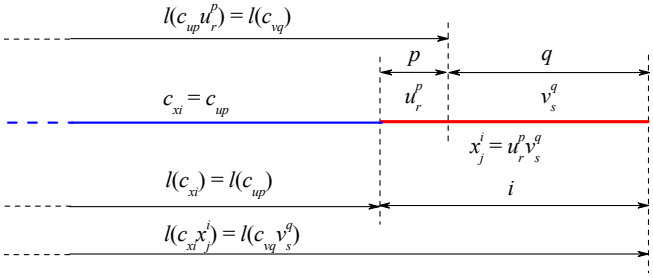


Figure 4. Meaningful composite word/sentence $x_j^i = u_r^p v_s^q$; v_s^q is a focal word of the sentence x_j^i, u_r^p is the focal word's fringe. Designations as in Figure 3.

As composite words are treated as BSDT PL sentences, the set of rules for the construction of meaningful composite words from a set of its possible constituents produces the BSDT PL *syntax*. Since any operations on simple and composite meaningful words should always be performed given their meanings (taking into account their common infinite beginnings), BSDT PL semantics (interpretations of words) is primary with respect to its syntax (rules for the construction of words) though they are of course closely related. If internal structure of meaningful composite words/sentences is ignored and they are perceived as a whole then communication with their help does not appeal to BSDT PL syntax and, consequently, it is performed *without syntax*, which is typical for animals and human infants [34].

3) Naming the things by meaningful words: All the meaningful words x_j^i constitute the BSDT PL *vocabulary* – a set of words that name, given their infinite context, all the things of the world, known as well as unknown but only conceivable. The number of things of the world is supposed to be infinite but countable, like the number of different meaningful strings $c_{xi}x_j^i$ related to a given proper class (Section VI A). BSDT PL vocabulary is always limited though, by request, may arbitrary be enlarged to the extent constrained mainly by particular animal's morphology only (new meaningful names may always be *constructed* and added to the vocabulary). Meanings of BSDT PL words are the ones that animals keep *actually* in their minds because, for an animal's survival, it is needed, its nervous system does not lie to itself. That is the reason why the BSDT PL should be successful as a truly primary language.

BSDT PL word x_j^i (i-bits-in-length sequence of +1s and –1s) is the jth pattern of spikes in the ith brain area equipped by the ijth BSDT ASM devoted to recognize the x_j^i, the name of the ijth thing of the world. This area may contain up to 2^i of such ASMs (cf. Figure 8). The reservation concerning brain areas (perceptual submodalities) is needed to connect the x_j^i to its context c_{xi} that specifies together with the x_j^i itself particular real-brain physical device recognizing the x_j^i and giving a meaning to it. Hence, BSDT PL meaningful words (patterns of +1s and –1s) are simultaneously the patterns of nerve impulses (spikes) in specific brain areas but certainly not the words of any of natural languages. With respect to our primary language natural languages are the *secondary* ones [26]. Invoking the notions of neuroscience (e.g., spikes or brain areas) for underpinning the theory's formal mathematical issues seems rather strange but does not reduce the theory's rigor. A reference to neuroscience is inevitably needed to give an explicit specification of one of the theory's principal paradigm shifts [36] shown by arrows in Figure 2, namely the shift from the domain of mathematical symbolism (box 4) to the domain of physically constructible brain devices (box 1) devoted to the recognition and processing of symbolically presented messages originated from things of the world (cf. Section X B and D).

C. Meaning Complexity and Levels of Meaning Uncertainty of Meaningful Words

The first thing that is needed to operate with meaningful words is a way of comparing them. Available methods do not hold in the case of concurrently infinite words.

1) The quest for a new infinity measure: All BSDT PL meaningful strings are one-way infinite, have an infinite length $\aleph_0$, and share their infinite beginning the length of which is again $\aleph_0$. Since their beginnings (distributed but precisely arranged and fixed "points of origin") are always the same (completely coincide), their end-points may take different locations and respective one-way infinite strings may in general be a number of bits longer or shorter with respect to each other. Figure 3 shows meaningful words whose complete string representations have the same end-points but whose string contexts have different end-points. Figure 5 illustrates another case: meaningful words whose complete string representations have different end-points but whose string contexts have the same end-points. We see the strings of the same infinite length in the sense of Cantor (that are countable) may be of different infinite length in the sense of the BSDT PL and, consequently, it is needed to

introduce a measure of lengths of such infinite strings (i.e., a measure of infinity) that should quantify their total lengths and the distinctions in positions of their end-points.

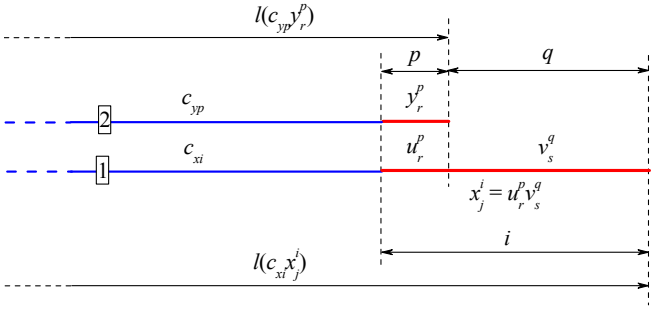


Figure 5. Meaningful words y_r^p and x_j^i of different in q bits meaning complexities, $l(c_{xi}x_j^i) - l(c_{up}y_r^p) = q$. Taken separately, strings 1 and 2 are of the same infinite length, $l(c_{xi}x_j^i) = l(c_{yp}y_r^p) = \aleph_0$. x_j^i and y_r^p are focal fractions of strings 1 and 2 and, consequently, are meaningful. If x_j^i is a composite word, $x_j^i = u_r^p v_s^q$, then u_r^p is a fringe of meaningful focal name v_s^q (in line 1 and 2, y_r^p and u_r^p may bit-by-bit coincide). Designations as in Figure 3.

2) Meaning complexity: We refer to the lengths $l(c_{xi}x_j^i)$ of infinite-on-a-semi-axis binary strings $c_{xi}x_j^i$ with their common infinite beginning (the context, c_{xi}) and their different explicitly specified affixes/meaningful *simple* words x_j^i as *meaning complexities* of these words (see Figures 3 to 5). If such strings have the same end-points then their affixes are understood as meaningful simple words of the same meaning complexity (Figure 3). If the end-points of such strings do not coincide then meaning complexities of respective meaningful words do not coincide too and the largest meaning complexity has the word whose meaningful string description has the end-point that is located to the right of end-points of other meaningful strings (in Figure 5, $c_{xi}x_j^i$ has larger meaning complexity than the $c_{yp}y_r^p$ because the difference $l(c_{xi}x_j^i) - l(c_{yp}y_r^p)$ equals $q > 0$). The length $l(c_{xi})$ of one-way infinite context string c_{xi} we call the *context complexity* of meaningful word x_j^i. For different meaningful simple words, their context complexities may be different (Figure 3) as well as the same (Figure 5). The length of the word x_j^i (it equals i bits) we call *ensemble* or *statistical* or *Boltzmann* or *Shannon complexity* of this word. To explain the latter, let us note that according to Claude Shannon [45], information or entropy of a set of 2^i of statistically independent binary messages x_j^i (the values of string variable x^i) is defined as $H(x^i) = -\sum_j \log_2(P(x_j^i))/|S_{xi}| = -2^i \log_2(1/2^i)/2^i = i$ where $P(x_j^i) = 1/|S_{xi}|$ and $|S_{xi}| = 2^i$ are respectively the probability of occurring of any of the x_j^i (they are here meaningless) and the total amount of different $x_j^i, j = 1, 2, \ldots, |S_{xi}|$.

Hence, meaning complexity of a meaningful simple word x_j^i equals the sum of its context complexity, $l(c_{xi})$, and its ensemble complexity, $l(x_j^i) = i$: $l(c_{xi}x_j^i) = l(c_{xi}) + i$ where the first item gives the length of complete description of the common part of the story of designing the devices devoted to the recognition of different meaningful words x_j^i and the second item gives the properties of any of the x_j^i averaged over the set of them, S_{xi}. String description $c_{xi}x_j^i$ of the story of designing the device devoted to recognize the x_j^i is actually *the shortest* evolutionary algorithm/instruction for such design and, consequently, the length of this algorithm/story is its *Kolmogorov* or *algorithmic complexity*, e.g., [46]. Of this follows, the notion of meaning complexity embraces the notions of Kolmogorov complexity/information and Shannon complexity/ information/entropy. Meaning complexity specifies the algorithm of designing a recognition device and reflects the complexity of this device dedicated to processing a particular meaningful word (an animal's respective internal/psychological state) and not the complexity of the thing named by this word.

3) Levels of meaning uncertainty: Let us now consider the case of meaningful *composite* words, e.g., $c_{xi}x_j^i$ with $x_j^i = u_r^p v_s^q$. If to *dynamically* fix p left-most components of an x_j^i as a particular u_r^p then $c_{xi}x_j^i = c_{xi}(u_r^p v_s^q) = (c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ where $c_{xi} = c_{up}$, $c_{vq} = c_{up}u_r^p$, $x_j^i \in S_{xi}$, $u_r^p \in S_{up}$, and $v_s^q \in S_{vq}$ (see line 1 in Figure 5). Infinite strings $c_{up}u_r^p \in S_{cu0}$ and $c_{vq}v_s^q \in S_{cv0}$ are the members of ultimate classes S_{cu0} and S_{cv0} that are different because they are generated by master strings c_{u0} and $c_{v0} = c_{x0}$ that share their beginning but differ in length in q bits. Owing to our infinity hypothesis the lengths of strings $c_{up}u_r^p$ and $c_{xi}x_j^i = c_{vq}v_s^q$ are comparable and the former is $l(c_{xi}x_j^i) - l(c_{up}u_r^p) = i - p = q > 0$ bits shorter (has smaller meaning complexity) than the latter (note, proper classes S_{cv0} and S_{cx0} coincide and, consequently, $l(c_{xi}x_j^i) = l(c_{vq}v_s^q)$). Since they both have infinite contexts, infinite strings $c_{xi}x_j^i$ and $c_{up}u_r^p$ are meaningful (in Figure 5, $c_{xi} = c_{up}$). But x_j^i (a simple focal word) and v_s^q (a focal fraction of $x_j^i = u_r^p v_s^q$) have the *definite* meanings while u_r^p (a fringe of the v_s^q in $x_j^i = u_r^p v_s^q$) a *conditional* meaning (see also Section IX). The latter may be compared with definite meanings of meaningful focal words with 2^q-state uncertainty defined by colored 2^q-non-locality of fringe words u_r^p. We refer to words whose meanings may only conditionally be defined as words of certain *levels of meaning uncertainty*. For this reason, the level of words of definite meanings is postulated to be zero (e.g., x_j^i, y_r^p, and v_s^q in Figure 5) while the level of uncertainty of words having conditional meanings (e.g., u_r^p in Figure 5) is a positive integer $q = i - p > 0$. The level of meaning uncertainty specifies the fringe's position in the body of its composite word (q bits to the left of the end-point of the whole meaningful string) and simultaneously, the degree, 2^q, of its colored non-locality. If there are words of definite meanings of different meaning complexity, e.g., x_j^i and y_r^p in Figure 5 then the one that has larger meaning complexity, x_j^i, may have a fringe, u_r^p, that coincides bit-by-bit with the word of definite meaning of smaller meaning complexity, y_r^p. In spite of that the meanings of y_r^p and u_r^p are essentially different. Attempts of comparing definite meanings of words of different meaning complexities (e.g., y_r^p and v_s^q or x_j^i) also lead to meaning uncertainties we quantify by the level of uncertainty of the u_r^p coinciding bit-by-bit with the y_r^p, i.e., the q for the example in Figure 5 (see Section IX).

4) Relative measurements of infinity using meaning complexity and levels of meaning uncertainty: Meaning complexity and levels of meaning uncertainty of BSDT PL meaningful words are the parameters needed to ensure a

relative comparison of lengths of one-way infinite strings sharing their infinite beginning and, as a result, the comparison of word meanings, definite as well as conditional. Traditional (in the sense of Cantor) comparison of lengths of such strings by counting the total amount of their bits has no sense here because resulting lengths are always the same and equal to $\aleph_0$ bits. Consequently, meaning complexity and levels of meaning uncertainty (as parameters specifying the infinity) exist in the framework of the BSDT PL only and have their roots in the hypothesis of concurrent infinity and the technique of proper classes. Meaning complexity embraces, given the context c_{xi}, Shannon-type ensemble complexity (the length of a word x_j^i in bits) specifying the word's ensemble properties (*averaged* over the ensemble of 2^i of x_j^i) and Kolmogorov-type algorithmic complexity (the length in bits of complete irreducible infinite evolutionary algorithm/instruction $c_{xi}x_j^i$ for designing the ASM that selects the meaningful x_j^i) specifying *the sameness* or *individual properties* of the device selecting the x_j^i and, through it only, *the sameness* or *individual properties* of the thing named by the x_j^i.

In this article, our meaning complexity is not compared with numerous other complexity definitions (the notion of the level of meaning uncertainty is new at all). We note only that most of them, to take into account the current actual context, attempt to estimate it, in one or another way, in a *finite* manner. For example, using a finite estimation of what is called an "effective complexity" (that is a loose counterpart to or a finite estimation of our context complexity), Murray Gell-Mann and Seth Lloyd [47] combine Kolmogorov complexity/information and Shannon complexity/information into a finite "total information." Hence, the meaning complexity's crucial distinction is *the genuine explicit infinity* of its descriptions of meaningful words/sentences – the faculty that is fundamentally impossible within the framework of standard ZFC or ZFC-like mathematics.

D. *Categories and Subcategories (Ontologies, Hierarchies) of Meaningful Words, Semantic Rule of Identity, Randomness and Irreducibility of Synonyms*

Meaningful words could be organized in structures that are themselves meaningful and have rich properties.

1) Categories and subcategories: The values $c_{xi}x_j^i$ of the form $C(x^i) = c_{xi}x^i$ define *a category* (notion, concept) of 2^i of meaningful words x_j^i that are here called *synonyms*. Meanings of these synonyms are given by the strings $c_{xi}x_j^i$, $x_j^i \in S_{xi}$ and $|S_{xi}| = 2^i$. Considering the x^i as a composite variable, $x^i = u^p v^q$, allows (given the context c_{xi}) a sub-categorization of items of the category $C(x^i) = C(u^p v^q)$. If to fix the context $c_{xi} = c_{up}$ and a value of u^p, e.g., u_r^p ($u_r^p \in S_{up}$ and $|S_{up}| = 2^p$) then we obtain the category's *the* pr*th subcategory* $C_{pr}(v^q) = (c_{up}u_r^p)v^q = c_{vq}v^q$ of synonyms v_s^q of definite meanings $(c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ where $v_s^q \in S_{vq}$ and $|S_{vq}| = 2^q$. If to fix the context $c_{xi} = c_{up}$ only then we obtain the category's the qth subcategory $C_q(u^p)$ of 2^p of synonyms u_r^p of conditional meanings of q-level meaning uncertainty (strings u_r^p occupy fringe positions in composite words $x_j^i = u_r^p v_s^q$, $q = i - p$). The number of subcategories $C_{pr}(v^q)$ whose synonyms have definite meanings equals the number of synonyms u_r^p of the subcategory $C_q(u^p)$ whose items have conditional meanings, i.e., 2^p; for the considered case of two-component composite words, the number of sub-categories whose synonyms have conditional meanings equals 1. If $p = 0$, the category $C(x^i) = C(u^p v^q)$ may be thought of as its own subcategory whose focal words $v_s^q = x_j^i$ ($q = i$) have common zero-length fringe u_0^0, $C(x^i) = C(u^p v^q) = C_{00}(v^q)$; meanings of synonyms of the $C_{00}(v^q)$ are given by the strings $c_{xi}x_j^i = c_{u0}u_0^0 v_s^q = c_{vq}v_s^q$ where $c_{xi} = c_{vq} = c_{u0}u_0^0$. For the case $i = 3$, Figure 6 demonstrates all possible sub-categorizations of the category $C(x^i) = C(u^p v^q)$: line 1 ($p = 0$, $q = 3$) and its fan of segments display $2^p = 1$ of subcategories $C_{00}(v^3)$ of $2^q = 8$ of items $(c_{u0}u_0^0)v_s^3 = c_{x3}x_j^3$ of definite meanings; line 2 ($p = 1$, $q = 2$) and its fan of segments show $2^p = 2$ of subcategories $C_{1r}(v^2)$ of $2^q = 4$ of items $(c_{u1}u_0^1)v_s^2$ of definite meanings and the only subcategory $C_2(u^1)$ of $2^p = 2$ of synonyms $c_{u1}u_r^1$ of conditional meanings of 2-level meaning uncertainty; line 3 ($p = 2$, $q = 1$) and its fan of segments depict $2^p = 4$ of subcategories $C_{2r}(v^1)$ of $2^q = 2$ of items $(c_{u2}u_0^2)v_s^1$ of definite meanings and the only subcategory $C_1(u^2)$ of $2^p = 4$ of synonyms $c_{u2}u_r^2$ of conditional meanings of 1-level meaning uncertainty. For the category $C(x^i) = C_{00}(v^i)$ and for all its subcategories $C_q(u^p)$ and $C_{pr}(v^q)$, the number of their zero-level words having definite meanings, regardless of the sort of their particular sub-categorization, always remains the same because $2^i = 2^{p+q}$ (in Figure 6, $2^i = 8$).

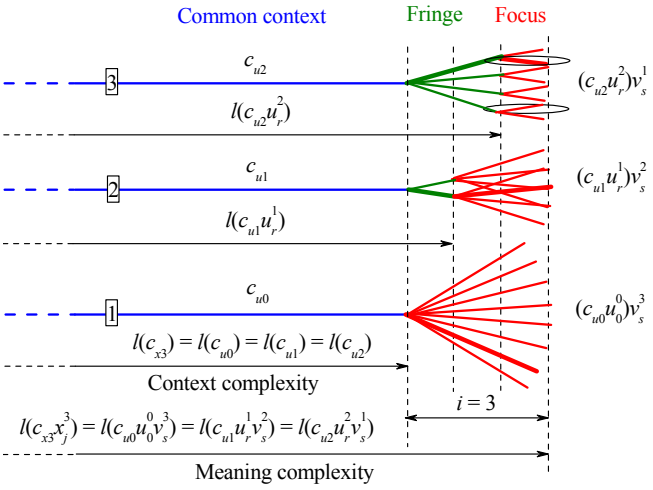


Figure 6. A category $C(x^i) = C(u^p v^q)$ of meaningful words and subcategories of them $C_q(u^p)$ and $C_{pr}(v^q)$, the case of $i = 3$. The largest common context, fringe and focal words are respectively shown as blue, olive and red line segments; thick line segments coincide completely bit-by-bit; definite meanings of circled words are directly incomparable, the distance between neighbor vertical dashed lines equals 1 bit. Other designations are as in Figure 3. See text for details.

Since synonyms are always defined given an infinite context, they should also be compared given the context. For example, in Figure 6, for subcategories $C_{00}(v^3)$, $C_{1r}(v^2)$ and $C_{2r}(v^1)$, their items shown as thick lines are bit-by-bit equivalent and the meanings of focal words $v_s^3 = x_j^3$, v_s^2, and v_s^1 of these strings may directly be compared. Since these focal words have different fringes or immediate contexts u_0^0,

u_r^2, u_r^1 and, consequently, different resulting total contexts $c_{u0}u_0^0$, $c_{u1}u_r^1$, $c_{u2}u_r^2$, they have different definite meanings and name the same thing's features (parts, attributes, properties, traits) under condition that "visual area" is gradually shrinking and covers gradually smaller number of "visible" features (from 8 for v_s^3 to 2 for v_s^1). Words that are circled in Figure 6 have different contexts and as a result directly incomparable meanings. Such a comparing should conditionally be performed.

2) Ontologies/conceptual spaces of meaningful words and hierarchies of brain devices for processing these words: We refer to particular arrangement of synonyms of definite and conditional meanings that are the members of all focal, $C_{pr}(v^q)$, and fringe, $C_q(u^p)$, subcategories of particular category, $C(x^i) = C(u^p v^q)$, as particular BSDT PL *partial* or *sub-ontology of meaningful words* that name the features of particular multi-feature thing of the world (in Figure 6 lines 1, 2, and 3 with their fans show three possible for this example sub-ontologies). Taken together sub-ontologies or conceptual subspaces devoted to meaningful naming the features of a particular multi-feature thing produce *the whole ontology or conceptual space* of words specifying this thing (all fans taken together in Figure 6). As has been mentioned earlier, in humans, the biologically plausible number of components of composite words is expected to be 3 to 5 [44] and, consequently, biologically plausible BSDT PL ontology is expected to contain 3 to 5 levels of subcategories. Zero-level subcategories represent the synonyms of definite meanings, all other the ontology's subcategories consist of items of conditional meanings with the level of meaning uncertainty q that has 3 to 5 grades of values. The case of more than two constituents of composite words as well the case of multiple multi-feature things will not be discussed.

Each synonym related to a given sub-ontology or ontology is processed in an animal's brain by a synonym-specific network/ASM device devoted to processing this synonym only. We refer to the arrangement of such real-brain recognition devices serving the sub-ontology or ontology as a real-brain network/ASM *sub-hierarchy* or *the whole hierarchy* (neural subspace or the whole neural space [28]) devoted to recognize and process the patterns of signals originated from different features of a multi-feature thing whose sub-ontology or ontology of names implements this network/ASM sub-hierarchy or hierarchy. Regardless of whether they are in focal or fringe positions, between the ontology's separate words (patterns of +1s and –1s), the hierarchy's separate recognition devices, and the separate features of a multi-feature thing of the world a one-to-one correspondence exists. It is assumed the hierarchy's recognition devices are real-brain implementations of BSDT ASMs (cf. Figure 8). The BSDT even makes a distinction between the ASMs dedicated to the recognition of fringe and focal constituents of composite words; they are *passive ASMs* and *active ASMs*, respectively [30]. As usually, the ontology's zero-level (focal) constituents have definite meanings (they are processed by active ASMs), all the other its constituents have conditional meanings and are processed by passive ASMs. This explains why we effortlessly recognize and name the whole multi-feature things (e.g., human faces) or their separate salient features (e.g., eyes or lips) and why we experience difficulties when recognizing and naming the relationships between these features or between these features and the whole thing. The same concerns the arguments for their allocation.

Multi-feature-thing-specific ontologies and network/ASM hierarchies may dynamically be constructed for temporal purposes in the process of an animal's adaptation to permanently changing environment (e.g., a short-term memory for the traffic on a street cross) and may almost "for ever" be embedded ("hardwired") into an animal's anatomy in the process of animal evolution, development and, finally, learning from experience (e.g., long-term memory for faces).

3) Semantic rule of identity: Given the context, the category/subcategory's synonyms name different sub-devices of the same multi-purpose real-brain recognition device and simultaneously respective features of the multi-feature thing of the world generated the signals this device is devoted to process. For this reason, direct *interchangeability* of synonyms is only possible among the members of the same category/subcategory (Figure 6): the change of a synonym changes the choice of an animal's focus of attention (Sections V B2 and VI B2) and changes the feature of current interest of *the given* multi-feature thing of the world – that is the BSDT PL's semantic *rule of identity*. In the BSDT PL, there is in principle no possibility of ascribing different names to the same thing or to the same feature of this thing because by definition each meaningful word is unique and its meaning is actually keeping in the mind of behaving organism. If it is not the case a malfunctioning of the organism appears.

4) Randomness and irreducibility of synonyms: To name a feature of a thing by one of 2^i of the category's synonyms x_j^i from the set S_{xi} (to teach one of 2^i of the hierarchy's sub-hierarchies to recognize the x_j^i), particular x_j^i (particular network/ASM sub-hierarchy devoted to store and recognize an i-bit message) is chosen *in random* because none of the category's features to be named, none of binary patterns x_j^i that may be used for their naming, and none of the network/ASM sub-hierarchies that may be chosen and taught to recognize the x_j^i have any priority over the others. Hence, for naming the features of a multi-feature thing, BSDT PL category's synonyms can only be chosen from their given range of values *in random*. But, on the other hand, once random choice of a name for naming a thing has been done (particular network/ASM sub-hierarchy has been taught to store and recognize a name), this name (respective sub-hierarchy) is then *rigidly* associated with one particular thing of the world or the thing's attribute. Thanks to this BSDT PL's peculiarity, it reconciles the notions of rigidity and contingency of names usually treated in semantics as different, e.g., [48].

The category's synonyms x_j^i can also be understood as *natural numbers* ranged from zero to $|S_{xi}| - 1 = 2^i - 1$ or from $2^i - 1$ to $2(2^i - 1)$ and written in binary string notations (see (1) in Section VII A). Since the synonyms constituting a category/subcategory are *randomly chosen natural*

numbers, they cannot be reduced to simpler mathematical expressions and are consequently *irreducible* (cf. Figure 8).

VI. BSDT PL Real Numbers, Continuum, and Continuity-discreteness Unity and Uncertainty

BSDT PL meaningful words could further be interpreted as traditional mathematical structures but written in a non-traditional way. This entails essential consequences.

A. Real Numbers and the Cantor's Continuum Hypothesis

BSDT PL meaningful words are traditional real numbers but have a common infinite beginning.

1) Real numbers and their countable continuum: All the strings $c_{xi}x_j^i$, generated by the same master string c_{x0}, have a common infinite beginning of the same infinite length $\aleph_0$ and taken together they produce the proper class S_{cx0}, $c_{xi}x_j^i \in S_{cx0}$. The number of elements of the ith fraction of the S_{cx0} (it comprises all the $c_{xi}x_j^i$ with affixes x_j^i that are not longer than i bits) is given by the sum $|S_{cx0}{}^i| = \sum 2^k = 2^{i+1} - 1$ where $|S_{cx0}{}^i|$ is the cardinality of the fraction $S_{cx0}{}^i$ and $k = 0, 1, \ldots, i$. Consequently, between natural numbers in their usual order and all the members of the S_{cx0} a one-to-one correspondence can be established (see also (1) and Section VII A). That means the cardinality of the S_{cx0}, $|S_{cx0}|$, and the cardinality of the totality of natural numbers, $\aleph_0$, are equal to each other, i.e., $|S_{cx0}| = \aleph_0$. On the other hand, if to posit $i = \aleph_0$ and neglect the 1s in above expression for the $|S_{cx0}{}^i|$ (they are in this case inessential) then it gives the cardinality of the S_{cx0} as $2^{\aleph_0}$, i.e., $|S_{cx0}| = 2^{\aleph_0}$ where $2^{\aleph_0}$ is the size of the Cantor's continuum.

The members of the proper class S_{cx0} do not exhaust the totality of all the BSDT PL's possible meaningful strings. Along with the master string c_{x0}, any of its infinite fractions c_{xI} that shares with the c_{x0} its infinite beginning but is shorter in I bits also generates its own proper class S_{cxI} of the size $\aleph_0 = 2^{\aleph_0}$ ($I = 1, 2, 3$ and so on without an upper limit). Each proper class S_{cxI} with $I > 0$ has the same number of items as the S_{cx0} (S_{cxI} with $I = 0$) but comprises meaningful words that have I-bits-smaller meaning complexity with respect to meaningful strings of the S_{cx0}. Hence, the totality of BSDT PL meaningful strings consists of all the members of all proper classes S_{cxI} that produce together *BSDT PL continuum – the totality of all its real numbers* (to remind, BSDT PL meaningful binary strings $c_{xi}x_j^i$ of the length of $\aleph_0$ bits can be understood as real numbers written as one-way infinite strings with common infinite beginnings and i-bits-in-length explicitly specified right-most fractions). As the number of elements of each of the classes S_{cxI} with $I = 0, 1, 2, \ldots$ is $\aleph_0 = 2^{\aleph_0}$ and the number of such classes is $\aleph_0$, the size of the totality of BSDT PL real numbers is also $\aleph_0 = 2^{\aleph_0}$. In other words, *BSDT PL continuum is countable*.

2) Refutation of Cantor's continuum hypothesis: The countability of the BSDT PL continuum radically contradicts to Cantor's continuum theory stated the existence of *two* kinds of infinities: the countable infinity of natural numbers (the cardinality of their totality is $\aleph_0$) and the uncountable infinity of real numbers (the cardinality of their totality is $2^{\aleph_0}$). With the help of its diagonal argument Cantor found that $\aleph_0 < 2^{\aleph_0}$. Additionally he conjectured his *continuum hypothesis*, *CH*, stating that there is no infinite set whose cardinality would be in between $\aleph_0$ and $2^{\aleph_0}$. But thanks to Kurt Gödel [49] and Paul Cohen [50], [51], it is known the CH is independent on the ZF or ZFC and can be neither proved nor disproved assuming the ZF/ZFC holds. It means simultaneously such extensions of ZF/ZFC are possible for which the CH either holds or fails. BSDT PL extension of the ZF/ZFC by the hypothesis of concurrent infinity is in this respect special because it leads to the *countable* continuum. This property of the continuum contradicts to Cantor's diagonal argument but is highly desirable of the view of such mathematicians as, e.g., Leopold Kronecker, Henri Poincaré, L. E. J. Brouwer or Hermann Weyl who were the opponents of Cantor's continuity/infinity theory. Thus, BSDT PL rather refutes than solves the CH or the first David Hilbert's problem [52].

Why in the case of the BSDT PL Cantor's diagonal argument does not work can be explained in the following way. We assume, as Cantor did, all the real numbers preexist and may be presented as infinite symbolic strings of the length $\aleph_0$. Cantor also postulated that each symbol in these strings is known and at any moment its identity, if one desires, may immediately be disclosed at least in principle. But contrary to Cantor, for each real number, we assume that the amount of its string symbols whose identities are known and at any moment may immediately be disclosed is always finite, whereas the amount of symbols of unknown identity is infinite. The reason for this fundamental distinction is that, in string descriptions of BSDT PL real numbers, the amount of their explicitly known symbols is always *finite* because they are defined by an always *finite* process of their construction.

B. Continuity-discreteness and other Related Unities and Uncertainties, Elusiveness of the Focus of Attention, the Symbolism's Insufficiency

Meaningful words in the form of real numbers sharing their infinite beginning provide a new view of possible symbolic representations of knowledge and computations.

1) Continuity-discreteness, quality-quantity, and reality-symbolism unities and uncertainties: The constructability of BSDT PL real numbers and the countability of the totality of them do not exert any influence on practical computations because they are in fact always performed with infinite real numbers presented by finite binary strings only. But the very fact of knowing the real numbers' constructability is of great practical importance because it shows that though the amount of symbolic information about a thing is always *infinite* its *finite* part may only be known at any moment.

A particular meaningful pattern of i signals originated from a thing of the world is processed by an animal's recognition device (an implementation of the BSDT ASM [30]) deliberately designed for this aim. The complete description of this device and the meaning of the pattern of signals it recognizes are given by an infinite binary string $c_{xi}x_j^i$ whose finite i-bits-in-length fraction x_j^i represents the pattern to be recognized. Consequently, *the finite* fraction of the $c_{xi}x_j^i$, x_j^i, does represent currently relevant symbolic,

explicitly formalized, quantitative, discrete-valued information about the thing. The remaining *infinite* fraction of the $c_{xi}x_j^i$ and *infinite* amount of symbolic information it contains, c_{xi}, are presented in a non-symbolic, informal, qualitative, implicit, real-valued, continuous way as the mentioned real-world recognition device. Hence, complete representation of any finite symbolic *meaningful* message should always consist of this message itself and of the real-world device devoted to recognizing it (cf. Section X B and D). The reason why we invoke such *unity or complementation of symbols and things* (in other cases the items of incommensurable domains) is our infinity hypothesis and our phenomenology equating infinite strings of symbols and real-world things (boxes 1 and 4, Figure 2).

Due to inherent unity of BSDT PL domains of symbols and of real-world things/real-brain devices, any *meaningful* discrete-valued description of any thing should inevitably contain some traces of continuity caused by intrinsic connections between meaningful patterns of symbols (finite vectors x_j^i) and real-brain things/devices (infinite strings $c_{xi}x_j^i$) devoted to process them. Purely discrete-valued symbolic messages are always meaningless, as in the case of Shannon [45]. BSDT PL messages or natural numbers or finite symbolic strings are meaningful because they are always conditioned by real numbers or one-way infinite strings of symbols. Of this follows a *continuity-discreteness* or *quality-quantity* or *reality-symbolism unity and uncertainty* of BSDT PL meaningful finite-in-length symbolic messages (cf. Section X E). That is why the world's BSDT PL discrete (quantitative, symbolic) and continuous (qualitative, real-world) representations must compete and coexist at anytime and anywhere. It is indeed the fact in real brains where cooperative and simultaneously competitive discrete-continuity effects are ubiquitous (it is sufficed to recall, e.g., close relationships between spike and wave neuron activities [53], [54]). Research concerning brain mechanisms and brain organization of uncertainty are reviewed, e.g., in [55]. An example is in Section X E.

2) Elusiveness of the focus of attention: In the domain of symbols, information exchange is performed by finite symbolic meaningless messages to be processed by Turing methods. In the domain of real-world devices, communication is performed by unspecified (e.g., chemical [21]) non-symbolic messages that can not be processed by Turing methods. Isolated symbolic messages take their meanings from their interactions with the domain of real-world devices. Such interactions define for an animal/ human the sort and the amount of currently relevant meaningful symbolic information to be communicated. Simultaneously, they define the sort and amount of non-symbolic but potentially symbolic contextual information that is not communicated together with symbols but is crucial for giving them meanings. The mechanism of selecting the relevant symbolic information (e.g., right-most line segments designated the whole words x_j^i or their "focal" fractions v_s^q in Figures 3 to 6) resides in the domain of things and defines for a perceiving agent/living organism the choice of its current *focus of attention*. Recent studies indeed demonstrate the importance of non-symbolic, e.g., wave-like interactions observed by EEG (electroencephalogram) or/and fMRI (functional magnetic resonance imaging) methods that are essentially continuous and implement large-scale network interactions supporting attention mechanisms in humans, e.g., [56], [57]. Moreover, a current empirical finding seems to directly indicate [58] that the purely symbolic approach, restricted to considering the spike brain activity only, is insufficient to explain the effects of attention and, consequently, "other processes must have a key role" [58], [59]. The reason is that an "SC inactivation caused major deficits in visual attention tasks" while simultaneous "attention-related effects in MT and MST remain intact," despite the usual view of selective spike activity of neurons in MT and MST as the main correlate and distinctive feature of visual attention [58]. SC, MT, and MST are respectively the superior collicus, middle temporal, and medial superior temporal monkey brain areas involved in motion-detection tasks.

3) Incompleteness of the BSDT PL formalism: BSDT PL discrete-valued formalism informs nothing of mechanisms of the arrangement of its composite words or sentences (Section V B2) and of selecting their fragments that are to be placed into the current focus of an animal's attention (Figure 4). Thus, in spite of its perfection and efficacy (Sections X and XI), discrete part of BSDT PL formalism is *incomplete* and, consequently, *insufficient* to ensure its own running in full. Its incompleteness is a manifestation of the continuity-discreteness or symbolism-reality uncertainty predicted by the BSDT PL. It is this uncertainty that is the reason why complete *symbolic* theory of anything, including the BSDT PL itself, is fundamentally impossible.

VII. BSDT PL Arithmetization by Natural Numbers and its Essential Randomness

Considering the right-most finite fraction of meaningful words as natural numbers provides unexpected solutions to some mathematical problems of great generality.

A. Arithmetization by Natural Numbers of Mathematical Expressions of Different Meaning Complexity

Every $c_{xi}x_j^i \in S_{cx0}$ generated by master string c_{x0} is uniquely labeled by its affix x_j^i or by its indices i and j. The x_j^i encodes i and j as its own length and its own *content* understood as an arrangement of this vector's positive and negative components. As it is usually done in computer sciences (see item two in (1)), strings x_j^i may also be treated as natural numbers written in binary notations and ranged from zero to $2^i - 1$. In such a form, the x_j^i with different i but same j correspond to same natural numbers. To ensure the unique bijection from strings to numbers, let us introduce decimal equivalents to strings x_j^i, the numbers G_{ij}^{x0}, as

$$G_{ij}^{x0} = \sum_{k=1}^{i} 2^{k-1} + \sum_{k=0}^{i-1} (x_j^i(k) + 1)2^{k-1} \qquad (1)$$

where $x_j^i(k)$ is the kth component of the x_j^i (it equals either +1 or –1); the second item of the sum (1) is, for this x_j^i, the value of j given the value of i. At $i \geq 1$, G_{ij}^{x0} = 1, 2, 3 and so on without an upper limit.

The affixes x_j^i of all the conceivable strings $c_{xi}x_j^i \in S_{cx0}$ may be treated as all the conceivable written in binary notations meaningful (given the c_{xi}) or meaningless (if the c_{xi} is ignored) mathematical expressions/assertions not longer than i bits. It means, strings x_j^i provide the labeling of themselves and of the mentioned expressions and consequently may be considered as *Gödel vectors/strings* that, because of (1), are one-to-one related to *BSDT PL Gödel numbers* G_{ij}^{x0}. Gödel numbers (1) are natural numbers and enumerate themselves and all the conceivable mathematical meaningful expressions $c_{xi}x_j^i$ or meaningless expressions x_j^i. Consequently, the numbers G_{ij}^{x0} provide for these expressions their complete *arithmetization by natural numbers*.

The G_{ij}^{x0} enumerate all the members of a given proper class, S_{cx0}. But the totality of such classes generated by different master strings c_{xI} is infinite (Section VI A) and for each of them, S_{cxI}, its own system of Gödel numbers G_{ij}^{xI} can by analogy be defined (I = 0, 1, 2, ... ; if I = 0, $G_{ij}^{xI} = G_{ij}^{x0}$). At different values of I, the totalities of strings x_j^i and natural numbers G_{ij}^{xI} are the same but they enumerate meaningful mathematical expressions of different meaning complexities (Section V C). The arithmetization just introduced is a non-Gödelian one though already in 1946 Kurt Gödel seemed to have envisaged something similar when he said about the possibility "to take the ordinals themselves as primitive terms" [60].

B. Natural Numbers, Gödel Numbers, Omega Numbers and Essential Randomness of their Use as Names

Every $c_{xi}x_j^i \in S_{cx0}$ contains an infinite-on-a-semi-axis and *always unspecified* initial part c_{xi} and a finite fraction x_j^i that, given the value of i, has *random* (incompressible and incomputable, Section V D4) arrangement of its binary components. This property of BSDT PL meaningful strings or *real numbers* (Section VI A) reflects their constructability ("random computable enumerability" [61]) and, given the value of i, the randomness of the x_j^i (its "algorithmic randomness" [62]). The x_j^i is to be processed by a special-purpose *self-delimiting computer* existing, we suppose, as a BSDT ASM [30] and dealing with random binary computer algorithms not longer than i bits. This property of BSDT PL words reflects their inherent connections with the devices that process them in the best way, of course, if their previous learning was perfect [31]. The mentioned properties of strings $c_{xi}x_j^i$ demonstrate that they are in fact binary string representations of *computably enumerable random real numbers* that, according to [61], are simultaneously *Omega-like* and *Omega numbers*. Gregory Chaitin's Omega number Ω gives the halting probability of randomly chosen binary algorithms running on a computer given its hardware and software [62]. The affix of the $c_{xi}x_j^i$, x_j^i, is an i-length fraction of Ω or a "partial" Ω, Ω_{ij}^{x0}, providing the halting probability of random binary algorithms running on the ijth self-delimiting in i bits computer. Omega-like numbers were introduced as a generalization of Ω but it later turned out, they and another generalization of Ω known as enumerable random real numbers are equivalent to Ω numbers [61].

The ijth numerically written partial halting probability can be presented [62] as

$$\Omega_{ij}^{x0} = \sum_{k=0}^{i} (x_j^i(k) + 1) / 2^{k+1} \qquad (2)$$

where x_j^i is the same probability written in binary string notations, $x_j^i(k)$ is the kth component of the x_j^i (it equals either +1 or –1), k is the length of a random binary algorithm running on the ijth computer (we suppose, BSDT ASM [30]), $1/2^k$ is the probability that this algorithm halts on this computer (if it halts, $(x_j^i(k) + 1)/2 = 1$ otherwise $(x_j^i(k) + 1)/2 = 0$). Halting probabilities Ω_{ij}^{x0} correspond to $c_{xi}x_j^i$ generated by the master string c_{x0}. Master strings c_{xI} whose meaning complexities are I bits smaller than the meaning complexity of the c_{x0} generate proper classes S_{cxI} and partial halting probabilities Ω_{ij}^{xI} with I = 0, 1, 2, … (if I = 0, $S_{cxI} = S_{cx0}$ and $\Omega_{ij}^{xI} = \Omega_{ij}^{x0}$; cf. Section VI A). At different values of I the totalities of strings x_j^i and halting probabilities Ω_{ij}^{xI} are the same but concern randomly chosen meaningful algorithms of different meaning complexity.

BSDT PL words x_j^i are *simultaneously* natural numbers, Gödel numbers, random algorithms, and partial Ω written in binary notations. That is, BSDT PL Gödel numbers G_{ij}^{xI} are essentially *random-valued*. This "strange" property can be explained if one notices that given the i the values of their indices j are randomly chosen from the range of zero to $2^i - 1$ or (see (1) and Section V C4) from $2^i - 1$ to $2(2^i - 1)$ where $2^i - 1 = G_{ij}^{x0}$ at $j = 0$ and $2(2^i - 1) = G_{ij}^{x0}$ at $j = 2^i - 1$. In other words, from 2^i of equal in rights ways of the enumeration of indices j, the second item in (1) gives only the one that was *randomly* chosen here for the reason of its analytical convenience only. As the size of the totality of different x_j^i is equal to the size of the totality of natural numbers $\aleph_0$ (Section VI A), the totalities of BSDT PL Gödel numbers (1), Chaitin numbers (2), random binary algorithms and self-delimiting computers devoted to process them are also of the size $\aleph_0$. As it was first demonstrated by Alan Turing [7], halting probabilities are incomputable. According to [62], they may be treated as true, unprovable assertions or irreducible mathematical facts (axioms) that in our case represent/name the things of the world or their particular features.

The totality of meaningful strings $c_{xi}x_j^i$ may be interpreted "as an alphabet of human thought" with the help of which "everything could be described and distinguished by means of the combination of the letters of this alphabet," to use words of Gottfried Leibniz (quotations from [63, p. 56]).

VIII. BSDT PL THRUTH AND UNDERSTANDING THE TRUTH

Defining and confirming the truths of meaningful words is inevitably needed for their successful practical use.

A. Convention on Truth

The name x_j^i is true if the truth value $T(c_{xi}x_j^i)$ of its meaning $M(x_j^i) = c_{xi}x_j^i$ is "true" or, in other words, if strings c_{xi} and x_j^i are correctly joined to each other. If there is no such correct correspondence, the truth value $T(c_{xi}x_j^i)$ of meaningful name is "false".

B. Completeness of Truths and Gödel's Incompleteness

Since the cardinality of the S_{cx0}, $|S_{cx0}| = \aleph_0$, is infinite, the number of BSDT PL truths is also potentially infinite and, for any meaningful string, its truth value $T(c_{xi}x_j^i)$ certainly exists and equals either "true" or "false". Each true meaningful name (infinite symbolic representation of a primary thought), e.g., $c_{xi}x_j^i$ names by definition the *ij*th real-world thing given to an animal through its *ij*th psychological state (physical implementation of the primary thought) or, in other words, through the activity of physically implemented real-world BSDT ASM devoted to process the x_j^i, ASM(x_j^i) [28], [30]. Thus, for meaningful words, the truth is the norm and the falsity is an anomaly caused, e.g., by an animal's dysfunction or disease. In any case, there is *no* lie and *no* liar paradox – a source of Kurt Gödel's incompleteness [6] which does not hold for BSDT PL *meaningful* [30], [32] names $c_{xi}x_j^i$ (Section VII). Axioms, theorems, and metamathematical expressions/assertions of a formal axiomatic system for which the Gödel's incompleteness holds are in our terms an infinite fraction of infinite in number *meaningless* strings x_j^i [17], [32]. These inferences are caused by the fact that BSDT PL name meanings are always the ones that animals/humans keep *actually* in their mind, like meanings of hypothesized by W. V. Quine "eternal" sentences [64]. It means, to survive, an animal does not lie to itself and it is the reason why the BSDT PL works so well as a primary language or a language of primary thoughts [28], [65], [66]. At the same time, a zero-level name's fringe words, due to their colored non-locality, have no definite but *conditional* meanings (Section IX and Figure 7). Fringe words are fuzzy/vague in meaning or "in limbo" in words of Quine [64], but their truths are *not* conditioned and always remain of certain values. The vagueness of meanings of fringe names is a BSDT PL manifestation, for the case of infinite sequences, of the famous *Burali-Forti paradox* (Section IX C) but it does not concern the truths.

BSDT PL convention on truth essentially differs from Alfred Tarski's *convention T* [67]. Tarski's definition is in fact *syntactical* and holds for an axiomatically defined pair object-language/meta-language only, whereas BSDT PL definition is *semantical* and uses the real world for checking the truths. Truth values of BSDT PL names are *unique* and *conclusive*. Any hierarchy of these truths is neither possible nor required, contrary to Tarski's syntactical approach implying that for any meta-language its meta-meta-language can in turn be conceived and so up. For this reason, for each of Tarski's meta-languages its higher-level meta-meta-language and respective higher-level truth (the truth of sentences of respective higher-level meta-language) could in general always be defined. Here, in words of Quine, "there is interlocking of class hierarchy with truth hierarchy" [64, p. 90], which may be traced back to early Bertrand Russell's theory of types [68]. Hence, Tarski's truths are *relative* while BSDT PL truths are *absolute*.

C. Discovering, Understanding and Confirming the Truths

The BSDT PL truths are here introduced as a *correspondence* between names (finite strings x_j^i) and things of the world (infinite strings c_{xi} or $c_{xi}x_j^i$). Indeed, each true meaningful name $c_{xi}x_j^i \in S_{cx0}$ names by definition the *ij*th real-world thing given to an animal through its *ij*th psychological state or the activity in the *ij*th BSDT ASM, ASM(x_j^i), designed, implemented in a physical form, and learned beforehand to recognize/select exactly the *ij*th thing by its symbolic name x_j^i [1], [17], [32]. As truth value $T(c_{xi}x_j^i)$ is never communicated together with the x_j^i, it should always be discovered in the process of *decoding* or *understanding* the meaning of the received name x_j^i and confirmed by checking the correspondence of this name to that reality or, more accurately, to an animal's psychological state (an activity of respective recognition device) represented this reality. In living organisms, it is most probably done by physical/anatomical segregation and specification of communication channels (input/output sensory submodalities) or/and by the choice of different physical carriers for different types of symbolic signals to be communicated. By means of such segregation and specification, the required *ij*th neural subspace/ASM hierarchy (the *ij*th computer for particular mental computations, Section VII B) is eventually allocated. By the following convergence of relevant channel-specific symbolic information from different communication channels an integral or holistic and, consequently, most reliable estimation of the current state of the animal's internal or/and external environment has to be achieved.

It is supposed the hierarchies (ontologies) of meaningful names/strings but *not* their truth values are implemented in the brain by means of BSDT neural subspaces/ASM hierarchies for signal processing, memory, decision-making and consciousness [28], [31], [32]. The truth value of each meaningful word of such hierarchy is not a property of the organism's device serving this word but a result of *evaluating* the state of this device. For this reason, of *the third person perspective,* it exists as a psychological state of an external observer who should intentionally define/ discover this value of truth ("true" or "false") by comparing the result of running the device serving the word of interest and respective thing of the world. Of *the first person perspective*, each meaningful word's truth value is postulated to be "true" because in terms of truths this fact reflects simply a distinctive feature of the definition of such words, namely that the word's meaning is the one that an animal actually stores in its mind.

IX. BSDT PL Meaning Ambiguity

Meaningful words are always the members of a proper class and this exerts essential influence on the possibility of comparing their meanings. In particular, in many cases, meanings are directly incomparable and, consequently, meaning ambiguities are inevitable.

A. Definite Meanings of Simple Words

It is assumed that, in a meaningful string $c_{xi}x_j^i$, its context c_{xi} and its simple focal name x_j^i describe respectively the *static* part or a "hardware" of the ASM(x_j^i) already fixed in the course of evolution and its *dynamic* part or "software" designed in the course of the hardware's adaptive learning

and development. The length of x_j^i in bits, i, defines the number of now essential (explicitly considered) features of the ijth thing named by the x_j^i; the jth arrangement of ±1 components of x_j^i is the jth BSDT PL description of this ijth thing (e.g., the value +1 or –1 of a component of the x_j^i may mean that the respective feature is included to, +1, or excluded from, –1, the consideration). The complexity of meaning of the name x_j^i reflects the meaning complexity of the physically implemented real-world ASM(x_j^i) and an organism of which the ASM(x_j^i) is a part but not the complexity of the thing named by x_j^i.

B. *Definite and Conditional Meanings of Constituents of Composite Words*

Different constituents of composite words are the members of different proper classes or of different meaning complexity. For this reason, relations between their meanings may be rather intricate.

1) "Virtual" devices for processing the "virtual" things: If the string x_j^i is a composite one, $x_j^i = u_r^p v_s^q$, then strings $c_{up}u_r^p$ and $(c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ describe, given the context $c_{xi} = c_{up}$, an ASM(u_r^p) and ASM(v_s^q) that may for a time period dynamically be created from the ASM(x_j^i) that in turn is the product of a similar process described by the string $c_{xi}x_j^i$. ASM(u_r^p) and ASM(v_s^q) are "virtual" ASMs (i.e., temporally designed for) selecting the names u_r^p and v_s^q of the prth and the qsth "virtual" things (i.e., of temporally highlighted/ allocated fractions of the ijth composite thing named by its ijth composite name x_j^i). In other words, virtual ASMs highlight the prth and qsth "partial" meaningful fractions of the ijth description of the ijth thing (cf. Figure 6). Composite names essentially enrich the BSDT PL semantics but raise the problem of comparing the meanings of names selected by ASM(u_r^p), ASM(v_s^q), and ASM(x_j^i).

2) Comparing the meanings of whole composite words and their focal fractions: Zero-level names x_j^i and v_s^q (v_s^q is a part of the $x_j^i = u_r^p v_s^q$) name given the context *the same* thing in the same way but from different points of view defined by their contexts (static for x_j^i, c_{xi}, and in part dynamically created for v_s^q, $c_{vq} = c_{up}u_r^p$; Figure 7(a)). The v_s^q is selected under condition $c_{xi} = c_{up}$ (for x_j^i and v_s^q their common context is c_{xi}) by the ASM(v_s^q) that is "virtual" with respect to the ASM(x_j^i). Thus, the ASM(x_j^i) can temporally serve as the ASM(v_s^q) but in any case the same thing is under the consideration and the meaning of x_j^i, $M(x_j^i) = c_{xi}x_j^i$, and the meaning of v_s^q, $M(v_s^q) = (c_{up}u_r^p)v_s^q = c_{vq}v_s^q$, may unambiguously be related. As thick line segments in Figure 6 demonstrate, a $c_{vq}v_s^q$ is simply another realization of the $c_{xi}x_j^i$.

3) Comparing the meanings of whole composite words and their fringe fractions: If u_r^p is a q-level fringe of zero-level focal string v_s^q and they are the fractions of the $x_j^i = u_r^p v_s^q$ (Figure 7(a)) then u_r^p has *no* definite meaning (Section V B and C). But it could get a *conditional meaning* if one supposes that u_r^p is conditioned by the color of a colored zero-level name $u_r^p(color)$ selected by a respective q-stages-back-in-evolution ASM. If it is, uncolored zero-level names x_j^i in Figure 7(a) are unambiguously related to colored zero-level names $u_r^p(color)$ in Figure 7(b). Vectors $u_r^p(color)$ and u_r^p conditioned by one of the q colors *color* have conditional but certain meanings. But once colors are deleted (only uncolored strings are used in computations) the one-to-one correspondence between x_j^i and $u_r^p(color)$ disappears and, instead of it, we obtain 2^q-state uncertainty between the x_j^i and u_r^p and between the definite meaning of x_j^i and conditional meaning of u_r^p (Figure 7).

4) Comparing the meanings of words naming evolutionary predecessors and successors: If, given the context $c_{xi} = c_{up}$, names x_j^i and u_r^p (or y_r^p in Figure 5) are both of the level of zero, then their meanings are to be of *different* proper classes and should have *different* meaning complexities (to remind, meaning complexity of x_j^i is $l(c_{xi}x_j^i) - l(c_{up}u_r^p) = i - p = q$ bits larger than that of u_r^p; see Figures 4 and 5, Figure 7(a) and (c)). This means they describe *different* things from the same point of view or the same thing at *different* stages of its evolution. The names x_j^i (Figure 7(a)) and u_r^p (Figure 7(c)) are respectively selected by present-stage-of-evolution ASM(x_j^i) and q-stages-back-in-evolution ASM(u_r^p) and refer to animals of evolutionary different species. Meaningful string $c_{up}u_r^p$ and respective part of $c_{xi}x_j^i = (c_{up}u_r^p)v_s^q$ may coincide bit by bit but even in this case meanings of x_j^i and u_r^p may only *conditionally* be related to each other and 2^q additional conditions (strings v_s^q in Figure 7(a)) are required to uniquely establish their correspondence.

5) Graphical illustration of meaning ambiguities:

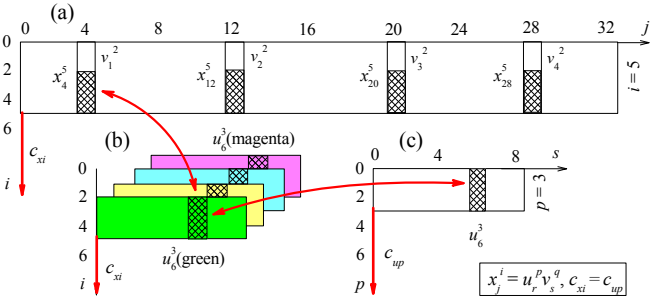


Figure 7. Comparing given the context different-level BSDT PL meaningful names of different proper classes: (a) zero-level names, (b) colored zero-level names corresponding to names in (a), (c) zero-level names that are predecessors to names in (a) and counterparts to names in (b).

In Figure 7, panel (a) demonstrates zero-level names x_j^i and v_s^q of meaningful strings $c_{xi}x_j^i$ and $(c_{xi}u_r^p)v_s^q$ ($x_j^i = u_r^p v_s^q$; $i = 5$, $p = 3$, and $q = i - p = 2$); the rectangle has the height i bits and the width $|S_{xi}| = 2^i = 32$ bits, the ijth bar of the height i in the jth horizontal position designates the name x_j^i that in a numerical form (see (1) and (2)) corresponds to Gödel number G_{ij}^{x0} and partial Chaitin number Ω_{ij}^{x0}; bars $x_4^5 = u_6^3 v_1^2$, $x_{12}^5 = u_6^3 v_2^2$, $x_{20}^5 = u_6^3 v_3^2$ and $x_{28}^5 = u_6^3 v_4^2$ that correspond to four colored highlighted bars in (b) are also highlighted (u_6^3 is q-level fringe of zero-level names v_s^q that is a focal fraction of the x_j^i); substrings v_1^2, v_2^2, v_3^2, and v_4^2 may encode the colors of colored strings $u_r^p(color)$ in (b). Panel (b) shows conditioned zero-level names $u_r^p(color)$ that in a numerical form (see (1) and (2)) correspond to colored Gödel numbers $G_{pr}^{u0}(color)$ and colored partial Chaitin numbers $\Omega_{pr}^{u0}(color)$; under condition that the word *color* is

a parameter, names u_r^p(*color*) are uniquely related to names x_j^i in (a) and selected by conditional q-stages-back-in-evolution ASMs; colored words u_r^p(*color*) conditionally name the things unconditionally named by the x_j^i; equal-in-size rectangles colored in $|S_{vq}| = 2^q = 4$ colors consist of $|S_{up}| = 2^p = 8$ bars of the height p; uncolored bars in (a) and respective colored bars in (b) (e.g., u_6^3(*green*) and x_4^5) denote different descriptions of the same thing. Panel (c) displays uncolored zero-level or focal names u_r^p (y_r^p in Figure 5) of meaningful strings $c_{up}u_r^p$ that name evolutionary predecessors of things named by the words x_j^i; the prth bar of the height p in the rth horizontal position (it is shaded) designates u_r^p for the case $u_r^p = u_6^3$ (in a numerical form it corresponds to Gödel number $G_{3,6}{}^{u0}$ and Chaitin number $\Omega_{3,6}{}^{u0}$).

In (a), (b), and (c), strings that are bit-by-bit equivalent to the u_6^3 are shaded in the same way. Contexts are shown as thick arrows and equal to each other bit by bit, $c_{xi} = c_{up}$. Uncolored and colored names name real-world unconditional and real-world conditional ("virtual") things, respectively. Between names in (a) and in (b) a bijection $x_j^i \leftrightarrow u_r^p$(*color*) exists that may be for example $x_{28}^5 \leftrightarrow u_6^3$(*magenta*) or $x_4^5 \leftrightarrow u_6^3$(*green*). A bijection also exists from names u_r^p in (c) to given-color names u_r^p(*color*) in (b). For example, it may be $u_r^p \leftrightarrow u_r^p$(*green*). But if such a bijection was already established then other conceivable bijections, e.g., $u_r^p \leftrightarrow u_r^p$(*magenta*) become impossible. Once colors are deleted, these bijections (they are indicated as curved bidirectional arrows) disappear producing, instead of 2^q-state (4-state in (b)) discrete colored non-locality of vectors u_r^p, 2^q-state (4-state in (b)) uncertainty (degeneracy) of meaning relations between names in (a) and (b), in (b) and (c), and in (a) and (c).

C. Relationships between Meaning Ambiguities and Burali-Forti Paradox

The origin of conditional relationships between meanings of names of different meaning complexities is the properties of ultimate/proper classes caused in turn by BSDT PL infinity hypothesis or vice versa, as the hypothesis of concurrent infinity was introduced when proper classes were already known in literature, e.g., [39]. Of this follows the famous Burali-Forti paradox according to which "there can be two transfinite (ordinal) numbers, a and b, such that a neither equal to, greater than, nor smaller than b" [8, p. 157] means in our terms that meanings of BSDT PL names whose meaning complexities differ in q bits can only be compared with 2^q-state uncertainty. In Figure 7 infinite strings $c_{xi}x_j^i$ and $c_{up}u_r^p$ are like Burali-Forti's transfinite ordinals a and b mentioned above.

The Burali-Forti paradox reflects the meaning-ambiguity properties of BSDT PL infinite symbolic statements/strings of different meaning complexities but in terms of transfinite ordinals. On the other hand, BSDT PL provides specific quantification of the ambiguities stated for different transfinite ordinals by the Furali-Forti paradox but in terms of BSDT PL strings of different meaning complexities. The reason is in end our infinity hypothesis.

X. Numerical and Empirical BSDT PL Validation

Now it is time to consider the BSDT PL validation.

A. Disappearing the Bounds between Mathematics and Reality

On the one hand, given the infinitely long context or "boundary conditions", the BSDT PL performs traditional mathematical computations with finite binary messages and is certainly a kind of mathematics that we call the mathematics of meaningful computations (Section IV). On the other hand, the BSDT PL is a kind of natural science because infinite-in-length boundary conditions used in its computations are implemented as real-world physical devices and, consequently, the computations themselves contain inevitably indispensable, inseparable from the symbolism elements of reality. Resulting *symbolism-reality unity and uncertainty*/*dichotomy* (cf. Section VI B) is not a failure or misunderstanding, it is the inherent property and distinctive feature of the BSDT PL caused directly by the hypothesis of concurrent infinity and its phenomenology formalization. Owing to this feature, within the BSDT PL framework, the distinctions between mathematics and reality, between mathematics and natural sciences become rather vague and sometimes disappear. That is also the reason why the BSDT PL can not be validated by the traditional in pure mathematics method of formal proofs, i.e., by deriving theorems from axioms. BSDT PL computations are conditioned by infinite-in-length context and for this reason contain some elements of mind/psychology (i.e., meanings of words) whereas standard mathematics ignores meanings by definition. That is why the only way to confirm the validity of the BSDT PL remains to compare its predictions with *real-world meaningful computations* that are abounds in living organisms. In sum, to validate the BSDT PL, it is needed to appeal to neuroscience, cognitive sciences, and psychology and compare their results with the BSDT PL predictions.

B. Solving the Communication Paradox

The first principal point needed to be understood is how in practice to communicate the meaningful words if they are by definition fundamentally infinite and how animals/humans solve this problem, routinely and immediately.

1) Serving the subconscious, basic behaviors and the simplest sociality: In Section V we saw BSDT PL simple words are those BSDT PL sentences that are perceived "holophrasically" and do have definite meanings. Internal structure of such sentences (the manifold of their possible focal and fringe constituents) is ignored and, consequently, they are presented *without BSDT PL syntax*. This fact and the fact that meanings of BSDT PL names are the ones that animals/humans keep *actually* in mind [1], [17], [30], [32] make the BSDT PL an appropriate tool for the description of communication without syntax or without any language at all – the style of communication that is typical for animals and human infants, e.g., [34] and references therein.

Meanings of BSDT PL words are simultaneously animal/human primary thoughts [32], i.e., the simplest or primitive or elementary patterns of involuntary, automatic

or *sub-* or *unconscious* activity of their brains. This activity is in fact the activity in a particular BSDT neural subspace/ASM subhierarchy (Sections V D and X E) that may most naturally be observed by an external observer as involuntary, automatic or *unconscious* behaviors of an animal that is under examination. We refer to these behaviors or body movements as basic or inherent ones because they truly reflect respective animal/human internal states. Among *basic behaviors* or invariant elements of a "paralanguage" [18] there are the ones (e.g., breathing or heart beating) that are truly innate and the ones (e.g., walking or directing the gaze) that originate from innate/ primitive behavioral reflexes, e.g., [69], [70], [71] but demand, after an animal's birth, for their further tuning and maturation in the course of "prepared" animal leaning and development (cf. Section XI E).

As the BSDT PL is well suited to describe not only meanings and primary thoughts but also basic behaviors, it is also capable of describing the behavioristic part [72] of an animal's cognition and based on basic animal behaviors communication without syntax or without any language at all. We hypothesize, this simplest type of communication suffices to support *the simplest* sociality that is typical in animals and human infants, e.g., [34].

2) Communication paradox: Since complete symbolic descriptions of BSDT PL meaningful names, $c_{xi}x_j^i$, are fundamentally *infinite*, during any finite time period none of them can ever be communicated in full even in principle while in fact many times a day everybody observes in others and experiences himself/herself numerous successful meaningful information exchanges. To cope with infinite symbolic messages for a finite time period, super-Turing devices with super-Turing computational capabilities are certainly required. Hence, this communication paradox [1], [17], [32] demonstrates that, in spite of the fact that real-world super-Turing computers are unknown and many experts believe even impossible [38], an everyday, routine, ubiquitous use of super-Turing computations is a norm in human meaningful communication.

3) Mirror transmitter and receiver devices for solving the communication paradox: In practice, communication paradox can be solved by appealing to BSDT infinity hypothesis (Section III and [1], [17], [32]) and the technique of BSDT ASMs [30]. On the one hand, these ASMs are devoted to process infinite-in-length meaningful messages but, on the other hand, they are special-purpose Turing computers running in the specific to each of them real-world environment. As in ASMs their programmatic and computational processes are in time completely separated, they do not waste their computational resources on serving themselves and, as any other special-purpose Turing computer, are faster than universal Turing computers [7].

But, dividing in time the programming and program running is insufficient to overcome the communication paradox. To cope with it, let us additionally suppose that in a communication process the ASM-transmitter and the ASM-receiver share in full their evolutionary history, i.e., they were designed, implemented in a physical form, and learned beforehand to perform the same meaningful function – selecting the same finite binary message x_j^i given the same infinite context or the same boundary conditions c_{xi}. If it is, and not in any other case, the meaning of x_j^i, $c_{xi}x_j^i$, is equally encoded, decoded, interpreted and *understood* by both parties and for both parties the value of its truth, $T(c_{xi}x_j^i)$, is the same. For this reason, and because the name's meaning is simultaneously a psychological state an animal experiences producing as well as perceiving this name, in the process of meaningful symbolic information exchange, the transmitter and the receiver are to be physically, structurally, and functionally *equivalent* in full or to be the "mirror" replicas or "clones" of each other (cf. Section X D). The fact that any two animal/human individuals, even identical twins or clones, always have different life-long individual experiences and, consequently, are never completely equivalent, is compensated by the tolerance of ASMs to their partial internal distortions and external noise [30], [73].

Several important BSDT PL predictions that are amenable for their empirical examination come out.

C. Coding by Synaptic Assemblies

Where meanings are essential (e.g., in living organisms) BSDT network learning paradigm "one-memory-trace-per-one-network" [28], [73] must be widespread in practice and, in particular, any memory for meaningful records must be built of the number of networks that coincides with the number of records to be stored in memory. This paradigm is not consistent with the usual desire of designers and engineers to store in a network as many traces as possible but it is the mandatory BSDT PL requirement ensuring the meaningfulness of memory records (Section V B).

A recent empirical neuroscience finding of coding by synaptic assemblies [74], [75] demonstrates this BSDT PL requirement is fulfilled in practice. In laboratory, mice were trained to perform new motor tasks. In behaving animals, changes in the number of synaptic contacts associated with learning new skills were measured. In complete accordance with the BSDT PL assumption that each new memory trace should be written down in an always new separate network or *synaptic assembly*, it turned out "that leaning new motor tasks (and acquiring new sensory experiences) is associated with the formation of new sets of persistent synaptic connections in motor (and sensory)" brain areas [76, p. 859].

D. Real-brain Mirror Neurons for Super-Turing Computations by Mirror ASMs

To ensure correct understanding of the meanings of finite symbolic messages, the ASM-transmitter and ASM-receiver that are the mirror replicas of each other are to be used (Section X B). These mirror ASMs implement *meaningful super-Turing computations*: for the transmitter and the receiver, they ensure the use of the same infinitely long "boundary conditions" c_{xi} needed to perform the previously programmed Turing computations with finite-in-length strings x_j^i, e.g., as in [28], [65], [66]. Mirror ASMs *physically* divide the infinite meaningful message to be processed into infinite, c_{xi}, and finite, x_j^i, parts and take the former into account as common for both parties "hardware",

designed and *physically* implemented beforehand in the course of animal evolution and development. Thanks to this "trick", to correctly understand the meaning of the $c_{xi}x_j^i$, it is enough to correctly transmit, receive, and decode the x_j^i only. The origin and theoretical substantiation of this trick is the BSDT PL phenomenology formalization (Sections III and IV) that equates infinite symbolic strings (e.g., c_{xi} or $c_{xi}x_j^i$) with real-world physical things.

Mirror ASMs also explain why meaningful communication without syntax is successful only between animals of the same or relative species: such animals are *a priori* equipped with the same "hardware" and "software" needed to finish meaningful super-Turing computations for a finite time period (small ASM changes or distortions do not matter because of ASM tolerance to damages and noise [30], [73]). The picture described is well supported by the empirical finding and studying of *mirror neurons* – the ones that are active when an animal behaves or only observes respective behaviors of others; see, e.g., [77], [78], [79] and numerous references therein. The mirror-ASM computational system just described and the mirror-neuron circuitries already observed in animals and humans [77], [78], [79] may respectively be treated as theoretical and real-brain implementations of until now hypothetical super-Turing machines with infinite inputs [38] that are capable of computing with infinite strings, which is the same as real-valued numbers.

One would object that the scheme proposed is nothing more than a regular Turing computer because nobody saw in animals anything else than regular Turing computations. But these Turing computations are "the tip-of-the-iceberg" of genuine super-Turing computations, the overwhelming part of which remains invisible if one looks for *symbolic* computations only. In the brain, conventional free-of-meaning Turing computations are immediately transformed into meaningful super-Turing computations once a finite symbolic message to be processed becomes rigidly connected (as it is actually the fact, Section X C) to its infinitely long context, *non-symbolically* presented as a real-world physical recall/recognition device or brain circuitry. In other words, any super-Turing computer processing a meaningful symbolic message might indeed be considered as a special-purpose regular Turing computer but running in the unique, specific to it *real-world environment that is also a part of computational process and computational device*. A Turing-type computer is inevitably a part of a super-Turing computer that is a *qualitatively distinct* computational machine because it combines symbolism and physical reality to process particular infinite-in-length symbolic strings or real-valued or continuous numbers for a finite time period. Another essential innovation is the use of *mirror* super-Turing machines, in order that communicators would be able to understand (correctly decode) a finite *meaningful* symbolic message addressed from one of them to another (see also Section X B3). BSDT PL super-Turing computations are not purely symbolic "tautological" transformations and super-Turing computers are not a set of connected elementary discrete-logic devices for doing these transformations – *both of them* are "an inseparable mix" of symbolism and reality. It is worth noting, that is also the reason why referring to this combined symbolism-reality computational method, we prefer the terms "primary language" and "language of primary thoughts" over "semantic mathematics" and "mathematics of meaningful computations."

E. Memory Performance without Knowing Memory Records, Continuity-discreteness Unity and Uncertainty

Since the BSDT PL employs for naming the things to be named a non-Gödelian arithmetization by natural numbers x_k^i (Section VII A) and since these natural numbers are *randomly* chosen from their finite range of values (Section V D4), particular random choice of a name does not matter and the following effect was predicted [17]. By empirical examination of an ASM hierarchy/neural subspace [28], [73] that generates the meaning of a trace x_k^i (Figure 8), all the parameters describing the ASM(x_k^i) may successfully be found but the content of the x_k^i – specific given the i randomly-established arrangement of its ±1 components – will always remain unknown. If it is, then, for example, the content of a particular given-length memory record does not affect memory performance and *can not empirically be found*. This rather surprising prediction [17] has well been corroborated by numerical BSDT PL analysis [66] of empirical receiver operating characteristics, ROCs (functions providing memory performance).

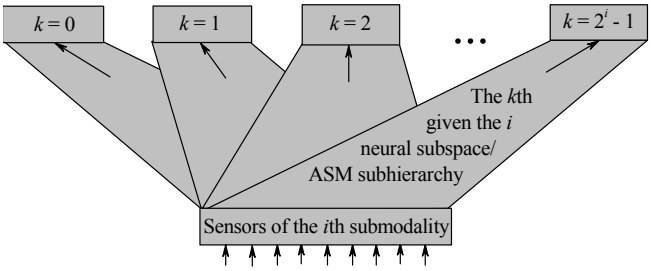


Figure 8. Neural subspaces/ASM subhierarchies generated the meanings of 2^i of given the i words x_k^i. All the subhierarchies (trapeziums) of the ith submodality are fed by impulses generated by the same set of sensors (lower rectangle) and the kth subhierarchy produces the kth pattern (arrow) of impulses to the kth apex ASM (ASM(x_k^i), the kth upper rectangle) learned to store and recognize the x_k^i. The correspondence between the kth given the i name (the arrangement of components of the x_k^i stored in the kth upper rectangle) and the thing it names is fixed but *randomly* established and empirically can not be found [66].

In [66] discrete-valued memory-for-meaningful-words ROCs measured in healthy humans and patients with brain disorders [80] were fitted by the BSDT. For this purpose, words and networks storing these words were presented as binary vectors x_k^i and respective BSDT ASMs. These ASMs are devoted to recall/remember/recognize the only memory trace specific to it [30]; for example, the ASM(x_k^i) serves the x_k^i. In Figure 8, the ASM(x_k^i) is an apex ASM of the kth ASM subhierarchy/neuron subspace generating the inputs to this ASM and giving the x_k^i its meaning (this scheme is called a semi-representational memory model [28], [73]). The size N of the network storing the x_k^i (in Figure 8, $N = i$), the intensity q of the cue used in the process of a memory

trace retrieval, the preferred rate j of the expected in experiment decision confidence, and the arrangements of components of vectors x_k^i (in Figure 8, they are enumerated by the index k) were used as fitting parameters. As a result of fitting, empirical discrete-valued ROCs for healthy subjects and patients with brain disorders were numerically reproduced by BSDT calculations (Figure 3 in [66]) and values of parameters N, q, and j were successfully found (Table 1 in [66]) *without* any reference to arrangements of components of vectors x_k^i. Thus, in full accordance with the BSDT PL prediction, complete BSDT description of performance of the memory-for-meaningful-words can indeed be achieved without knowing the memory records. Consequently, the ideas of non-Gödelian BSDT PL arithmetization by natural numbers (Section VII A) and, simultaneously, BSDT semi-representational memory model [28], [73] have indeed numerically and empirically been substantiated.

In spite of the BSDT's essential discreteness, it contains one *continuous* physical/physiological parameter, namely the neuron triggering threshold θ [66], [81]. In the course of ROC fitting [66], all the values of the θ from their jth finite-in-width range $\Delta\theta_j$, $\theta \in \Delta\theta_j$, are transformed into the only *integer* value of the decision confidence j from its range $0 \le j \le N + 2$. Reverse transformation of this value of j into a certain $\theta \in \Delta\theta_j$ that just generated the j is impossible and this fact is the BSDT implementation [66] of the predicted by the BSDT PL continuity-discreteness unity and uncertainty (Section VI B). This prediction and its BSDT implementation are well supported by the empirical discovery of irremovable spike onset potential (neuron triggering threshold) variability up to 10 mV [82]. In terms of neuroscience this variability is explained by fluctuating synaptic currents and by inherent statistics of the opening cell channels while in BSDT terms it is the width of the $\Delta\theta_j$ in voltage units.

XI. Examples and Perspective BSDT PL Applications

Practical examples of BSDT PL meaningful computations could help to better understand their features and the perspectives of their further applications.

A. BSDT PL Computations, the Concept and the Manual

The BSDT PL's *concept* of semantic computations is surprisingly simple because it recommends, before the beginning of calculations, *to know* complete formal description of reality. At the same time, it is surprisingly complex because it recommends, before the beginning of calculations, *to find* complete formal description of the reality. As such a description (the context) is by definition of an infinite length, these recommendations can of course never be fulfilled completely. That is, the main problem of meaningful computations is the incompleteness of knowledge of their context or, in other words, the incompleteness of available formal or "mathematical" descriptions of reality. As soon as such a description has been found and fixed, BSDT PL semantic computations are reduced to usual Turing computations and could easily be performed, e.g., [28], [65], [66].

In addition to these general recommendations, the BSDT PL gives also *a manual* for meaningful computations. It is based on the BSDT (a theory providing the best encoding-decoding rules [31, 73] for binary finite-dimensional vectors x_j^i damaged by replacing binary noise [27]) and the technique of BSDT ASMs (abstract selectional machines [30] implementing the BSDT encoding-decoding rules or BSDT PL inference rules). Given the context, the BSDT implements the main distinct features of meaningful BSDT PL computations: 1) the discreteness of all the computations with finite binary vectors x_j^i [65], [81], 2) the uniqueness of the vector x_j^i a particular ASM is devoted to process in the best way [30], and 3) the ability of each ASM to generalize even from a single example [73]. The first of these features leads to a fundamental discreteness of all BSDT PL computational predictions found at precisely fixed context. The second feature generates one-memory-trace-per-one-network network learning paradigm. The third feature ensures the BSDT PL's tolerance to damages and noise and its capability of coping with "effective stochasticity" of an agent's permanently changing environment.

For the study of meaningful information exchanges, their actual context should empirically be estimated with maximal possible accuracy. This is not a trivial problem and it is a subject of intensive research, e.g., [83], [84], [85]. Results available in this field are so far insufficiently rich because the required measurement methods remain till now in the state of development.

B. Meanings of Traditional Solutions of Mathematical Problems of Science and Practice

Besides the axioms, any formal axiomatic system, FAS, comprises symbolic descriptions of all its theorems and inference rules. In the BSDT PL, an FAS (e.g., ZFC) is represented as an *infinite fraction* of meaningless finite binary strings, x_j^i, that are the affixes of meaningful strings $c_{xi}x_j^i \in S_{cx0}$ (Section V B). For this reason, any FAS computations are also BSDT PL computations and numerous already available computational results, e.g., in physics or biology may be treated as examples of BSDT PL computations performed given a context defined *formally and informally*.

A separate infinite BSDT PL string $c_{xi}x_j^i$ that gives a meaning $M(x_j^i) = c_{xi}x_j^i$ to a finite symbolic message x_j^i (it may be, e.g., a physical formula written in binary notations) includes an infinite description c_{xi} of the FAS needed to derive this formula and of the physical problem that gives this formula a physical sense. For professionals (P) and laypersons (L), this formula has different meanings we denote as $M_P(x_j^i)$ and $M_L(x_j^i)$, respectively: $M_P(x_j^i) = c_{xi}(P)x_j^i = c_{xi}(IP)c_{xi}(FP)x_j^i$ and $M_L(x_j^i) = c_{xi}(L)x_j^i = c_{xi}(IL)c_{xi}(FL)x_j^i$ where finite-in-length strings $c_{xi}(FP)$ and $c_{xi}(FL)$ represent the formal knowledge (F) and infinite-in-length strings $c_{xi}(IP)$ and $c_{xi}(IL)$ represent informal knowledge (I) about the formula and the problem of interest. Formal knowledge

can be found in books or any other relevant texts, informal knowledge can only be acquired from individual experiences of professionals and laypersons, respectively.

Since professionals and laypersons have essentially different backgrounds in a particular knowledge domain, $c_{xi}(IP)c_{xi}(FP) \neq c_{xi}(IL)c_{xi}(FL)$, they understand the meaning of the x_j^i in a different way. Since $c_{xi}(FP)$ and $c_{xi}(FL)$ are finite and may explicitly be specified, they may be compared explicitly (e.g., by the grading of school exams). Since $c_{xi}(IP)$ and $c_{xi}(IL)$ are one-way infinite and essentially unspecified, it is impossible to compare them explicitly. We may only know that they are of different meaning complexities (Section V C) and have somewhere "in the past" a common infinite initial part. If to remember our phenomenology formalization (Figure 2) then it becomes clear that informal or "implicit" knowledge is presented in the brain as non-symbolical properties of real-brain devices serving this knowledge. Informal character of implicit knowledge also indicates the crucial role of the teacher and educational environment (the supervisor and research environment) for acquiring knowledge. Consequently, as an important source of informal knowledge, the teacher/ supervisor can never be excluded from the process of teaching/research training.

Formal or "explicit" knowledge – strings $c_{xi}(FP)$ and $c_{xi}(FL)$ – is not the content of books we have read in a school or university but their individual internal symbolic representation that, in terms of the BSDT PL, may be different in different minds. Informal or implicit knowledge – strings $c_{xi}(IP)$ and $c_{xi}(IL)$ – we acquire from our *personal* experiences under different teachers/supervisors in different educational/research environments and, consequently, it is also different for each of us. For these reasons, different professionals and different laypersons read the same books but understand them differently. In particular, for different professionals which we call P_1 and P_2, the formula x_j^i always has to an extent different meanings, $M_{P1}(x_j^i) \neq M_{P2}(x_j^i)$, and, consequently, even in so-called "exact" sciences a vagueness of meanings of their formal results is unavoidable and can not completely be excluded. In other words, for traditional FAS computations their context and, consequently, their meanings for different peoples can never *precisely* be fixed. Hence, in this case, the BSDT PL may only approximately be applied: its inherent discreteness is masked by the vagueness of knowledge each of us have about the context of ZFC computations. At the same time, as all humans (professionals and laypersons) are of the same species, their knowledge is internally represented by infinite strings of the same meaning complexity and, consequently, all of us can understand anything that understands anyone else on the condition of course that beforehand we were equally prepared/trained (see also Section X D).

Relationships, which were just described, between formal/informal knowledge and traditional computations, draw our attention to the fact that any manipulations with numbers will be meaningless until their giving-the-meaning context is added and fixed.

C. Processing Meaningful Memory Records and Meaningful Images Given their Precisely Fixed Contex

A situation may of course be conceived when the context of different symbolic messages is completely, bit-by-bit the same without any reservations. *For particular animal/human*, it may be, e.g., the case of members, $c_{xi}x_j^i$, of the same category of names, $C(x_j^i)$. If the context of names x_j^i, c_{xi}, is *precisely fixed* then the inherent discreteness of the BSDT PL should be visible as inherent discreteness of respective empirical data and these inherently discrete data should successfully be described by the inherently discrete BSDT PL computations. The main obstacle is the need of discovering such inherently discrete natural phenomena and developing a methodology of research that will not hide their discreteness.

As has been demonstrated in [28], [65], [66] all the mentioned conditions can be satisfied. As a result, we have already three particular *examples of complete successful application of given the context discrete-valued BSDT PL computations* to account for practically important cognitive (where the role of mind is essential) phenomena in humans. They are in particular 1) judgment errors in cluttered environments [65], 2) remembering/retrieving the words from a memory for meaningful words [66], and 3) recognition of meaningful images (human faces) by healthy humans [28]. In the first of these cases, with the help of the BSDT, the data measured in rating experiments when healthy subjects identify target stimuli in a cluttered visual environment, confound them with competing stimuli, and demonstrate high confidence of their erroneous decisions were quantitatively explained (Figure 4 in [65]); in the second case, memory-for-meaningful-words ROCs measured in healthy humans and patients with brain disorders were quantitatively described by the BSDT and memory-for-meaningful-words parameters were found (Figure 3 and Table 1 in [66]); in the third case, psychometric functions measured in human face recognition experiments were reproduced by the BSDT keeping the Neyman-Pearson objective (Figure 5 in [28]).

In each of these examples, BSDT discrete-valued numerical analysis has been applied to fitting empirical data measured by traditional techniques and analyzed by the authors of original publications [80], [86], [87] using traditional continuous computations. The authors of these publications did not recognize the discreteness of their results, in particular, because of the essential *continuity* of mathematical models they employed for empirical data analysis. We take the opposite view of these models motivated by the BSDT PL and its hypothesis of concurrent infinity. Namely, results of cognitive experiments found at the precisely defined context are to be *discrete* because in such a case the continuous/real-valued component of meaningful messages (infinite-in-length context or the set of real-brain circuits/devices involved in serving the cognitive tasks) is strictly the same, fixed, excluded from explicit consideration and "invisible" in practice as a result. Indeed, if a given person recalls different meaningful words or recognizes different meaningful images then he/she is

dealing with different finite symbolic messages x_j^i *given* bit-by-bit the same context c_{xi} defined by this person's unique previous experience that shapes his/her unique mind, relevant to a particular problem. This mind or particular set of real-brain learned circuits or particular neuron subspace or particular context c_{xi} is surely the same in all the tasks of the same type this mind (person) currently performs (cf. Figure 8). On the other hand, it is the discreteness of empirical data that is a manifestation of certainly definite meanings of symbolic messages, involved in context specific cognitive tasks. In other words, meaningfulness of messages to be processed and their precisely known context are simultaneously the source of this type of data. The continuity of observed cognitive performance indicates the vagueness of the context's estimation (inability of keeping the fixed context) that may be caused by irrelevant choice or inaccurate use of measurement protocols. It is supposed, if meanings are fixed and exactly communicated then the context is completely the same and communication (encoding/ decoding) performance should surely be discrete. As the fitting of empirical data measured in these three different types of cognitive experiments demonstrates [28], [65], [66], some popular study protocols seems to produce discrete-valued data in cognitive sciences though, to be convinced, new BSDT fitting results and some control experiments are surely required [66].

The need to keep the same context to ensure accurate meaningful communication/computation is rather well known, e.g., [83], [84], [85]. In cases where this demand is accurately satisfied, methods developed by other authors coincide with the BSDT PL sometimes almost literally. For example, what is called in [88] "meaning-generating capacity" of a complex dynamic system, namely "the proportion between the size m of the set of final attractor states and the size n of the set of all initial states of a system, i.e., $MC = m/n$" is in BSDT terms the ASM(x_k^i) probability of correct decoding given the size of the network $N = i$, intensity of cue q, and decision confidence rate j. This probability is denoted as $P(N, q, j)$ or $P(N, q, F_j)$ (F_j is false alarm probability given the j) and was already successfully used to analyze the results of real cognitive experiments [28], [65], [66]. The main distinction between the MC and the $P(N, q, j)$ is that the former is defined given the finitely estimated context whereas the latter is fixed given an infinitely defined context. The fact that the $P(N, q, j)$ is perfect [31], [89] and may even analytically be found [89] is secondary with respect to the context infinity.

D. A Lecturer and Students in a Lecture Room

Let us consider a lecturer who intends to deliver students the meaning of a physical formula as he/she understands it. At the beginning of a lecture, he/she and his/her students have different background knowledge of the formula of interest and students can not correctly understand its meaning. The lecturer's aim is to give them a piece of additional knowledge and, in this way, to equalize, for all of them, the context needed to equally understand the formula's meaning. At the end of the lecture, for the lecturer and for his/her students, infinite BSDT PL strings describing this specific knowledge should become bit-by-bit equivalent not only "in the past" but also "in the present", and the formula's meaning should be understood by all the parties in the same way. If it is not for any reason, a misunderstanding arises.

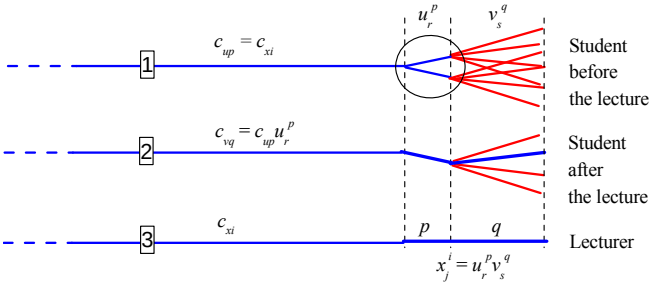


Figure 9. BSDT PL model of information exchange between the lecturer (line 3) and students before (line 1) and after (line 2) the lecture. The amount of information of interest is q bits, the "gap" (it is circled) between the knowledge of the lecturer and the knowledge of students equals p bits (examples for $q = 2$, $p = 1$; $i = q + p$). The fan of line segments attached to line 1 represents given the context c_{up} the set of possible meanings of the formula of interest in students before the lecture. The fan attached to line 2 gives the set of possible meanings after the lecture (after bridging the knowledge gap), thick line segments coincide bit-by-bit with the BSDT PL representation of the lecturer's knowledge (the thick fraction of the line 3). Other designations as in Figure 6.

In Figure 9 some aspects of the process of teaching and learning are presented in BSDT PL terms. It means, we ignore so far the fact that the lecturer and students use a natural language for their communication and we are not interested in mechanisms of translation of a natural language into the BSDT primary language and vice versa. As the lecturer and students are of the same species, they all use the same primary language (we suppose, BSDT PL) and their meaningful words are of the same meaning complexity. Such a representation is person-dependent and natural-language-independent. Line 1 in Figure 9 represents the ith submodality (a particular set of brain circuits that are ready to be changed or a plasticity area, cf. Figure 8) allocated in the brain of a student before the lecture to write, store and then retrieve particular information this student intends to acquire from the lecture. Line 2 gives the same for a successful student after the lecture; line 3 represents the lecturer's same brain area. One-way infinite fractions of these lines, $c_{xi} = c_{up}$ (from the left to the first vertical dashed line), designate a description of everything that is in common in the lecturer and his/her students, from genetic code to textbooks they have read. Composite vector $x_j^i = u_r^p v_s^q$ describes new knowledge the lecturer intends to deliver. This vector can be divided into two constituents one of which (v_s^q) describes the physical formula of interest and the other (u_r^p) describes the additional information ("local context") needed to connect the v_s^q to already available background knowledge, $c_{xi} = c_{up}$. A vector u_r^p or, more accurately, the form u^p (Section V B) may be treated as p-bits-in-width "gap" between the knowledge of the lecturer and the knowledge of students that should be bridged before the students would be able to understand the formula. As a

result of teaching, a student's area of plasticity responsible for acquiring this specific knowledge changes and reduces 2^i of different ways of understanding the message the lecturer communicates to the only one, the same as of the lecturer. It means, all the parties have now the same BSDT PL internal representation of the knowledge of current interest (thick line segments in lines 2 and 3) and equally understand it.

E. Non-syntactic and Non-language Communication by Basic Bihaviors

In the previous example (Section XI D), non-syntactic messages represent a very small fraction of the general flow of information. Among them it may be, e.g., the facts that the lecturer is walking when he/she gives his/her talk. This non-syntactic and even non-language message (bodily signal) is effortlessly understood by everyone who is in the room because all people are members of the same species, have the same *innate* bodily infrastructure and the same basic behaviors (Section X B1) developed from their same *innate* behavioral reflexes, e.g., [69], [70], [71] in the course of human learning and natural ageing (maturation) in a mostly common environment. As a result, adults or infants of relevant ages have *common mirror neuron systems* (Section X B3 and D) to produce and perceive/understand their basic behaviors included, for example, walking. For this reason, humans/animals produce and perceive such non-syntactic non-language messages originated from their basic behaviors automatically, with practically no chance of misunderstanding. For all given the species animals (or humans), meanings of their basic behaviors are given by practically equivalent brain circuits or infinite BSDT PL strings of the same length (to remind, possible the strings' distinctions are inessential because of BSDT ASM tolerance to damages and noise [30], [73], Section X B3).

The role of mirror neuron systems for understanding the actions and possibly the intentions of others, e.g., [90], [91], [92] and their role in evolutionary language development are rather well recognized [93], [94] and even to a degree studied by the method of computational modeling, e.g., [95], [96]. In these publications, the importance of training the innate brain structures for a design, on their ground, of mirror neuron systems and the importance of mirror systems for mimicking actions and language production are in particular emphasized. At the same time, contrary to the BSDT PL assumption, Michael Arbib and his colleagues suppose [93] - [96] that super-Turing computability is not relevant to brain computations. Such an attitude seems indeed rather natural while we are dealing, as it is usually the case, with *meaningless* computations only and do not pay attention to their meanings. But as soon as meanings become essential super-Turing computability escapes from the shadow of conventional Turing computations and becomes crucially important. It is what is the case for the BSDT PL because it does imply that super-Turing computability, as an indispensable part of animal/human communication process, maintains mechanisms of doing all the meaningful actions the brain serves (Section X B3 and D), including all kinds of non-language and language meaningful information exchange.

F. Natural Languages and Consciousness, Intuition, Free Will and Creativity

One of distinctive features of the BSDT PL is that truth values of its names are always in the norm true (Section VIII). That is why it serves so well to as a primary language for maintaining an animal's ongoing internal activity. For the same reason, it can serve as a "source language" whose meaningful words (an animal's psychological states) may next be translated into vocal, gesture, etc tokens of a more elaborate symbolic communication system needed to support information exchange between animals of a group. The more complicated the group's sociality, the more complicated communication system is required to support it, and vice versa. Since among other animals humans do have most complicated sociality, human natural languages are to be most complicate and elaborate. The BSDT PL may be used as a basis for the construction of such "secondary" [26] languages whose capacities may be up to the level of human natural language capacity. If so, semantics and syntax of natural languages should be based on semantics and syntax of the BSDT PL and should be implemented by mechanisms of (and innate brain structures for) the translation of words/sentences of the primary language into words/ sentences of a secondary language. In that sense the BSDT PL is a counterpart or a precursor to what is known as Noam Chomsky's "universal grammars," e.g., [97], [98].

The BSDT PL phenomenology formalization literary equates one-way infinite binary strings and animal psychological states or subjective experiences/qualia. It also represents given the context computations with finite binary meaningful strings as operations with animal subjective experiences. In other words, the BSDT PL solves the "hard" problem of consciousness [33] (the quest for a description of subjective experiences) as a whole and at once, simply by *postulating* the logically strict BSDT PL definition of qualia (Sections III and V B). For this reason, the BSDT PL is actually a theory of subjectivity (meaning, feeling, perception) or *a theory of the subconscious*. The problem remains to apply this theory to solving particular practical consciousness problems, as it has, e.g., been done in the case of BSDT atom of consciousness model, BSDT AOCM [32]. In particular, the mentioned above problem of translating the primary language into a secondary one may also be treated as the problem of translating the subconscious *served by super-Turing computations* into the conscious *served by Turing computations*. In both cases, the process of translation should inevitably be based on so far unspecified mechanisms of *intuition, free will* and *creativity*.

XII. CONCLUSIONS

The BSDT PL is based on the hypothesis of concurrent infinity and its phenomenology formalization (Sections I to IV). It provides what is called a "paradigm shift" [36]: a possibility to equate the items of such usually incommensurable domains as the symbolism and reality, to define strictly indefinable in traditional mathematics notions of meaning and subjectivity, and to perform explicitly given the context meaningful computations. Such computations

are an inherent mix of symbolism and specific to it reality implemented by a qualitatively novel computational device – a super-Turing computer with infinite inputs implemented in an animal's brain as a system of mirror neurons (Section X D). The range of perspective BSDT PL applications covers everything where meanings are important or, in other words, everything we, humans, may be interested in. In this sense it may be "a theory of everything". As soon as meanings become inessential, the BSDT PL is reduced to traditional ZFC mathematics. Available empirical and computational results support this view (Sections X and XI).

BSDT PL provides a framework that is sufficient to perform principal semantic computations and based on them communication without syntax. BSDT PL seems also to be sufficient to explain the computational part of intelligence of animals of poor sociality and, consequently, to design the computational part of intelligence of artificial devices (e.g., robots) or computer codes mimicking the behavior of such animals. At the same time, the BSDT PL is unable to symbolically explain the mechanism of splitting its composite words (sentences) into focal and fringe constituents (Section VI B2) and, consequently, of directing an animal's attention to a particular thing – we hope it may be done by methods beyond the discrete BSDT formalism. To explain/reproduce the "attentive" part of animal intelligence in a biologically-plausible way and to design the "attentive" part of the intelligence of intelligent robots, analog (e.g., wave-like) computational methods similar to those that are used in real brains are most probably required.

Contrary to traditional formal languages, e.g., [98], [99] that are in end the products of traditional ZFC mathematics, the BSDT PL is a consistent and complete (Sections VI to VIII) calculus of finite binary strings (spike patterns or "symbols") with *infinitely* defined contexts. It is based on 1) the new infinity hypothesis and its phenomenology formalization (Sections I to IV) providing the technique of super-Turing (semantic) computations with infinite binary strings that share their infinite initial part and 2) the BSDT [25] and its ASMs [30] providing a technique for the best encoding/decoding in binary finite-dimensional spaces [27] and implementing BSDT PL inference rules. BSDT PL is the simplest language of its kind and has great potential for designing the adequate models of higher-level languages, including in perspective the natural languages of humans. At the same time, meaning ambiguity of BSDT PL names of different meaning complexity that has been established as their fundamental property (Section IX) raises many intriguing problems to be solved in the future.

The BSDT PL describes a way for the communication of meanings of symbolic messages by means of basic animal behaviors (Sections X B1 and XI E) that could represent the behavioristic part [72] of more complex adaptive animal behaviors. For animals of the same and, in many cases, of relative species, thanks to their mirror neurons, e.g., [77] - [79] and common "bodily infrastructure" [34], it is intelligible without any efforts. For animals with most primitive sociality (including human infants) or for their *artificial* counterparts, a version of the discrete BSDT PL formalism may serve as an *exhaustive* but *incomplete* (Section VI B3) set of tools needed for their routine communication. How the primary language generates secondary (natural) languages and consciousness [32] is the problem of future research.

Following Rudolf Carnap [100, p. 204] let us finish this article by a quotation from Bertrand Russell (his term "denotation" may here be understood as "meaning"): "Of many other consequences of the view I have been advocated, I will say nothing. I will only beg the reader not to make up his mind against the view—as he might be tempted to do, on account of its apparently excessive complication—until he has attempted to construct a theory of his own on the subject of denotation. This attempt, I believe, will convince him that, whatever the true theory may be, it cannot have such a simplicity as one might have expected beforehand" [101, p. 518].

ACKNOWLEDGMENT

I thank the participants of the conference INTELLI-2012 for their comments on my oral presentation and Prof. Michael Arbib for his comments to its published version. I am grateful to my family and my friends, especially to my son Dr. Mykhaylo Gopych, for their help and support.

REFERENCES

[1] P. Gopych, "Primary language for semantic computations and communication without syntax," in INTELLI 2012 – The First International Conference on Intelligent Systems and Applications, P. Lorenz and P. Dini, Eds. Chamonix, France, 2012, pp. 47-53.

[2] G. H. Hardy. A mathematican's apology. Cambridge, England: Cambridge University Press, 1941.

[3] L. A. White, "The locus of mathematical reality: An apological footnote," in The World of Mathematics, vol. 4, J.R. Newman, Ed. New York: Simon and Shuster, 1956, pp. 2348-2364.

[4] A.A. Fraenkel, Y. Bar-Hillel, and A. Levy. Foundations of set theory, 2nd ed. Amsterdam: Elsevier, 1973.

[5] W. Sieg, "Hilbert's program: 1917-1922," Bull. Symb. Logic, vol. 5, March 1999, pp. 1-44, http://www.jstor.org/stable/421139.

[6] K. Gödel. On formally undecidable propositions of Principia Mathematica and related systems, English transl. by B. Meltzer. New York: Dover, 1992. Originally published Kurt Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," Monatshefte für Mathematik and Physik, vol. 38, 1931, pp. 173-198.

[7] A. Turing, "On computable numbers, with an application to the Entscheidungsproblem," Proc. London Math. Soc. (2) vol. 42, 1936, pp. 230-265; vol. 43, 1937, pp. 544-546; in Collected works of Alan Turing. Mathematical logic, R.O. Gandy and C.E.M. Yates, Eds. Amsterdam: Elsevier, 2001, pp, 18-54.

[8] H. Poincaré. Science and method. London: Thomas Nelson and Sons, 1914.

[9] S.C. Kleen. Introduction to metamathematics. Amsterdam: North-Holland Publ., 1952.

[10] A. Heyting. Intuitionism. An introduction. Amsterdam: Noth-Holland Publ., 1956.

[11] J.R. Lucas, "Minds, machines and Gödel," Philosophy, vol. 36, Apr.-Jun. 1961, pp. 112-127.

[12] R. Penrose. The emperor's new mind. New York - London: Pinguin Books, 1991.

[13] R. Penrose. Shadows of the mind: a search for the missing science of consciousness. Oxford: Oxford University Press, 1994.

[14] J.R. Searle. The mistery of consciousness. New York: New York Review Book, 1997.

[15] W.T. Fitch. The evolution of language. Cambridge, UK: Cambridge University Press, 2010.

[16] H. Weyl, Philosophy of mathematics and natural scienves, 4th ed. Prinecton: Princeton University Press, 1959.

[17] P. Gopych, "On semantics and syntax of the BSDT primary language," in KDS 2010, Information models of knowledge, K. Markov, V. Velychko, and O. Voloshin, Eds. Kiev-Sofia: ITHEA, 2010, pp. 135-145, http://foibg.com/ibs_isc/ibs-19/ibs-19-p15.pdf <retrieved: November, 2012>.

[18] E.O. Wilson. Conscilience: the unity of knowledge. New York: Vintage Books, 1999.

[19] H. Maturana and F. Varela. Autopoiesis and cognition. Dordrecht, Holland: D. Reidel Publ., 1980.

[20] L. Margulis. Symbiosis in cell evolution, 2nd ed. San Francisco: Freeman, 1993.

[21] C.B. Pert, M.R. Ruff, R.J. Weber, and M. Herkenham, "Neuropeptides and their receptors: a psychosomatic network," The Journal of Immunology, vol. 135 (Suppl.), Aug. 1985, pp. 820-826.

[22] M. Minky. The society of mind. New-York: Simon and Schuster, 1988.

[23] P.O.A. Haikonen. Robot brains: circuits and systems for conscious machines. Chichester, UK: John Wiley and Sons, 2007.

[24] D. Dennett. Consciousness Explained. London: Penguin Books, 1993.

[25] P.M. Gopych, "Elements of the binary signal detection theory, BSDT," in New research in neural networks, M. Yoshida and H. Sato, Eds. New York: Nova Science, 2008, pp. 55-63.

[26] J. von Neumann. The computer and the brain. New Haven: Yale University Press, 1956.

[27] P. Gopych, "BSDT multi-valued coding in discrete spaces," in CISIS-08, ASC, vol. 53, E. Corchado, R, Zunino, P. Gastaldo, and Á. Herrero, Eds. Berlin-Heidelberg: Springer, 2009, pp. 258-265, doi: 10.1007/978-3-540-88181-0_33.

[28] P. Gopych, "Biologically plausible BSDT recognition of complex images: the case of human faces," Int. J. Neural Systems, vol. 18, Dec. 2008, pp. 527-545, doi: 10.1142/S0129065708001762.

[29] G.M. Edelman, "Naturalizing consciousness: a theoretical framework," Proc. Natl. Acad. Sci., vol. 100, Apr. 2003, pp. 5520–5524, doi: 10.1073/pnas.0931349100.

[30] P. Gopych, "Minimal BSDT abstract selectional machines and their selectional and computational performance," in IDEAL-07, LNCS, vol. 4881, H. Yin, P. Tino, E. Corchado, W. Byrne, and X. Yao, Eds. Berlin-Heidelberg: Springer, 2007, pp. 198-208, doi:10.1007/978-3-540-77226-2_21.

[31] P.M. Gopych, "Foundations of the neural network assembly memory model," Int. J. Comput. Res., vol. 13, 2004, pp. 103-166; in Leading-edge computer sciences, S. Shannon, Ed. New York: Nova Science, 2006, pp. 21–84.

[32] P. Gopych, "BSDT atom of consciousness model: the unity and modularity of consciousness," in ICANN-09, LNCS, vol. 5769, C. Alippi, M.M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds. Berlin-Heidelberg: Springer, 2009, pp. 54-64, doi: 10.1007/978-3-642-04277-5_6.

[33] D. Chalmers. The conscious mind. Oxford: Oxford University Press, 1996.

[34] N.J. Enfield, "Without social context?" Science, vol. 329, Sep. 2010, pp. 1600-1601, doi: 10.1126/science.1194229.

[35] E. Husserl. The idea of phenomenology, Enlish transl. by L. Hardy. Dordrecht: Kluwer Academic Publ., 1907/1999.

[36] T.S. Kuhn. The structure of scientific revolutions, 3rd ed. Chicago-London: Universithy of Chicago Press, 1996.

[37] L. Wittgenstein. Tractatus logico-philosophicus. London - New York: Routledge, 1922/1974.

[38] T. Ord, "The many forms of hypercomputations," App. Math. Comp., vol. 178, Jul. 2006, pp. 143-153, doi: 10.1016/j.amc.2005.09.076.

[39] W.V. Quine. Set theory and its logic. Cambridge, MA: Harvard University Press, 1969.

[40] L. Pozsgay, "Liberal intuitivism as a basis of set theory," Proc. Sym. Pure Math., vol. 13, part I. Providence, Rhode Island: AMS, 1971, pp. 321-330.

[41] D. Perrin and J.-É. Pin. Infinite words. Amsterdam: Academic Press, 2004.

[42] K. Mainzer. Thinking in complexity, 5th ed. Berlin: Springer, 2007.

[43] V. McGee. Truth, vagueness, and paradox. Indianapolis: Hackett, 1991.

[44] S. Ullman, "Object recognition and segmentation by a fragment-based hierarchy," Trends Cogn. Sci., vol. 11, Feb. 2007, pp. 58-64, doi: 10.1016/j.tics.2006.11.009.

[45] C. Shannon, "A mathematical theory of communication," Bell Syst. Techn. J., vol. 27, 1948, pp. 379-423, 623-656.

[46] M. Li and P. Vitanyi. An introduction to Kolmogorov complexity and its applications, 2nd ed. Berlin-Heidelberg: Springer, 1997.

[47] M. Gell-Mann and S. Lloyd, "Information measures, effective complexity, and total information," Complexity, vol. 2, Sep.-Oct. 1996, pp. 44-52, doi: 10.1002/(SICI)1099-0526.

[48] S. Kripke. Naming and necessity, rvd. ed. Oxford: Basil Blackwell, 1990.

[49] K. Gödel. The consistency of the axiom of choice and of the generalized continuum hypothesis with the axioms of set theory. Princeton, NJ: Princeton University Press, 1940.

[50] P.J. Cohen, "The independence of the continuum hypothesis I," Proc. Natl. Acad. Sci. USA, vol. 50, Dec. 1963, pp. 1143-1148.

[51] P.J. Cohen, "The independence of the continuum hypothesis II," Proc. Natl. Acad. Sci. USA, vol. 51, Jan. 1964, pp. 105-110.

[52] D. Hilbert, "Mathematical problems. Lecture delivered before the International congress of mathematicians at Paris in 1900," English transl. by M. W. Newson, Bull. Amer. Math. Soc., vol. 8, 1902, pp. 437–479.

[53] K. Whittingstall and N.K. Logothetis, "Frequency-band coupling in surface EEG reflects spiking activity in monkey visual cortex," Neuron, vol. 64, Oct. 2009, pp. 281-289, doi: 10.1016/j.neuron.2009.08.016.

[54] B.S.W. Ng, N.K. Logothetis, and C. Kayser, "EEG pase patterns reflect the selectivity of neuron firing," Cerebral Cortex, Feb. 2012, doi: 10.1093/cercor/bhs031.

[55] D.R. Bach and R.J. Dolan, "Knowing how much you don't know: a neural organization of uncertainty estimates," Nat. Rev. Neurosci., vol. 13, Aug. 2012, pp. 572-586, doi: 10.1028/nrn3289.

[56] G. Thut, A. Nietzel, S.A. Brandt, and A. Pascual-Leone, "Alpha-band electroencephalographic activity over occipital cortex indexes visuospatial attention bias and predicts visual target detection," J. Neurosci., vol. 26, Sep. 2006, pp. 9494-9502, doi: 10.1523/jneurosci.0875-06.2006.

[57] T.P. Zanto, M.T. Rubens, A. Thangavel, and A. Gazzely, "Causal role of prefrontal cortex in top-dowm modulation of

visual processing in working memory," Nat. Neurosci., vol. 14, May 2011, pp. 656-661, doi: 10.1038/nn.2773.

[58] A. Zénon and R.J. Krauzlis, "Attention deficits without cortical neuronal circuits," Nature, vol. 489, Sep. 2012, pp. 434-437, doi: 10.1038/nature11497.

[59] A. Smolyanskaya and R.T. Born, "Attention is more than meets the eye," Nature, vol. 489, Sep. 2012, pp. 371-372, doi: 10.1038/489371a.

[60] K. Gödel, "Remarks before Princeton bicentennial conference on problems of mathematics," in The undecidable, M. Davis, Ed. New York: Raven Press, 1965, pp. 84-88.

[61] C.S. Calude and G.J. Chaitin, "Randomness everywhere," Nature, vol. 400, Jul. 1999, pp. 319-320, doi: 10.1038/22435.

[62] G. Chaitin. The limits of mathematics. Singapore: Springer, 1998.

[63] G. Lolli, "Peano and the foundations of arithmetic," in Giuseppe Peano between mathematics and logic, F. Skot, Ed. Milan: Springer, 2011, pp. 47-67.

[64] W.V. Quine. Pursuit of truth. Cambridge, MA: Harvard University Press, 1992.

[65] P.M. Gopych, "Performance of BSDT decoding algorithms based on locally damaged neural networks," In IDEAL-06, LNCS, vol. 4224, E. Corchado, H. Yin, V. Botti, C. Fyfe, Eds. Berlin-Heidelberg: Springer, 2006, pp. 199-206, doi: 10.1007/11875581_24.

[66] P. Gopych and I. Gopych, "BSDT ROC and cognitive learning hypothesis," in CISIS-10, AISC, vol. 85, Á. Herrero, E. Corchado, C. Redondo, Á. Alonso, Eds. Berlin-Heidelberg: Springer, 2010, pp. 13-23, doi:10.1007/978-3-642-16626-6_2.

[67] A. Tarski. Logic, semantics, metamathematics, 2nd ed. Oxford: Oxford University Press, 1935/1983.

[68] B. Russell. Principles of mathematics. London - New York: Routledge, 1903/2010.

[69] J.L. Pressler and J.T. Hepworth, "Newborn neurologic screening using NBAS reflexes," Neonatal network, vol. 16, Sep. 1997, pp. 33-46.

[70] J. Scott and M. Rosser, "The grasp and other primitive reflexes," J. Neurol. Neurosurg. Psychiatry, vol. 74, May 2003, pp. 558-560, doi: 10.1136/jnnp.74.5.558.

[71] L.E. Berk. Child development, 7th ed. Boston, MA: Allyn and Bacon, 2006.

[72] S.J. Shettleworth. Behavior, cognition, evolution, 2nd ed. Oxford: Oxford University Press, 2010.

[73] P.M. Gopych, "Generalization by computation through memory," Int. J. Inf. Theo. Appl., vol. 13, Apr.-Jun. 2006, pp. 145-157.

[74] T. Xu, X. Yu, A.J. Perlik, W.F. Tobin, J.A. Zweig, K. Tennant et al., "Rapid formation and selective stabilization of synapses for enduring motor memories," Nature, vol. 462, Dec. 2009, pp. 915-919, doi: 10.1038/nature08389.

[75] G. Yang, F. Pan, and W.-B. Gan, "Stably maintained dendritic spines are associated with lifelong memories," Nature, vol. 462, Dec. 2009, pp. 920-924, doi: 10.1038/nature08557.

[76] N.E. Ziv and E. Ahissar, "New tricks and old spines," Nature, vol. 462, Dec. 2009, pp. 859-861, doi: 10.1038/462859a.

[77] G. Rizzolatti and L. Craighero, "The mirror-neuron system," Ann. Rev. Neurosci., vol. 27, 2004, pp. 169-192, doi: 10.1146/annurev.neuro.27.070203.144230.

[78] C. Keysers and V. Gazzole, "Expanding the mirror: vicarious activity for actions, emotions, and sensations" Curr. Opin. Neurobiol., vol. 19, Dec. 2009, pp. 666-671, doi: 10.1016/j.conb.2009.10.006.

[79] C. Keysers, J.H. Kaas, and V. Gazzole, "Somatosensation in social perception," Nat. Rev. Neurosci., vol. 11, Jun. 2010, pp. 417-428, doi: 10.1038/nrn2833.

[80] A.P. Yonelinas, N.E. Kroll, J.R. Quamme, M.M. Lazzara, M.J. Sauve et al., "Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity," Nat. Neurosci., vol. 5, Oct. 2002, pp. 1236-1241.

[81] P.M. Gopych, "Sensitivity and bias within the binary signal detection theory, BSDT," Int. J. Inf. Theo. Appl., vol. 11 Oct.-Dec. 2004, pp. 318-328.

[82] B. Naundorf, F. Wolf, and M. Volgushev, "Unique features of action potential initiation in cortical neurons," Nature, vol. 440, Apr. 2006, pp. 1060-1063, doi: 10.1038/nature04610.

[83] J. Klüver and C. Klüver. On communication. An interdisciplinary and mathematical approach. Dortrecht, The Netherlands: Springer, 2007.

[84] M.C. Frank and N.D. Goodman, "Predicting pragmatic reasoning in language games," Science, vol. 336, May 2012, pp. 998, doi: 10.1126/science.1218633.

[85] C. Kemp and T. Regier, "Kinship categories across languages reflect general communicative principles," Science, vol. 336, May 2012, pp. 1049-1054, doi: 10.1126/science.1218811.

[86] S. Baldassi, N. Megna, and D. Burr, "Visual clutter causes high-magnitude errors," PloS Biol., vol. 4**,** 2006, e56.

[87] G. Rhodes and L. Jeffery, "Adaptive norm-based coding of facial identity, " Vision Res., vol. 46, Sep. 2006, pp. 2977–2987, doi: 10.1016/j.visres.2006.03.002.

[88] J. Klüver, "A mathematical theory of communication: meaning, information, and topology," Complexity, vol. 16, Jan.-Feb. 2011, pp. 10-26, doi: 10.1002/cplx.20317.

[89] P.M. Gopych, "ROC curves within the framework of neural network assembly memory model: some analytic results," Int. J. Inf. Theo. Appl., vol. 10**,** Apr.-Jun. 2003, pp. 189-197.

[90] V. Gallese and A. Goldman, "Mirror-neurons and the simulation theory of mind reading," Trends Cogn. Sci., vol. 2, Dec. 1998, pp. 493-501, doi: 10.1016/S1364-6613(98)01262-5.

[91] M. Iacoboni, "Neural mechanisms of imitation," Curr. Opin. Neurobiol., vol. 15, Dec. 2005, pp. 632-637 doi: 10.1016/j.conb.2005.10.010.

[92] J.M. Kilner, K.J. Friston, and C.D. Frith, "The mirror-neuron system: a Bayesian perspective," Neuroreport, vol. 16, Apr. 2007, pp. 619-623, doi: 10.1097/WNR.0b013e3281139ed0.

[93] M.A. Arbib, "From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics," Behav. Brain Sci., vol. 28, Apr. 2005, pp. 105-167.

[94] M.A. Arbib. How the brain got language: the mirror neural system hypothesis. Cambridge, MA: MIT Press, 2012.

[95] J.B. Bonaiuto, E. Rosta, and M.A. Arbib, "Extending the mirror neuron model, I: audible actions and invisible grasps," Biol. Cybern., vol. 96, Feb. 2007, pp. 9-39, doi: 10.1007/s00422-006-0110-8.

[96] J.B. Bonaiuto and M.A. Arbib, "Extending the mirror neuron model, II: what did I just do? A new role for mirror neurons," Biol. Cybern., vol. 102, Apr. 2010, pp. 341-359, doi: 10.1007/s00422-010-0371-0.

[97] N. Chomsky. The minimalist program. Cambridge, MA: MIT Press, 1997.

[98] W.J.M. Levelt. An introduction to the theory of formal languages and automata. Amsterdam-Philadelphia: John Benjamins Publ., 2008.

[99] J.H. Hopcroft and J.D. Ullman. Formal languages and their relations to automata. Reading, MA: Addison-Wisley, 1969.

[100] R. Carnap. Meaning and necessity. Chicago, Illinois: Chicago University Press, 1948.

[101] B. Russell, "On denoting," in Philosophy for the 21st century, S.M. Cahn, Ed. Oxford: Oxford University Press, 2003, pp. 512-518. Originally published B. Russell, "On denoting," Mind, 1905, vol. 14, pp. 479-493.

IEEE 802.11g Radio Coverage Study for Indoor Wireless Network Redesign

Sandra Sendra[1], Diana Bri[2], Emilio Granell[3] and Jaime Lloret[4]

Instituto de Investigación para la Gestión Integrada de zonas Costeras - Universidad Politécnica de Valencia, Spain

[1]sansenco@posgrado.upv.es, [2]diabrmo@upvnet.upv.es, [3]emgraro@posgrado.upv.es, [4]jlloret@dcom.upv.es

***Abstract*— An efficient wireless design and development is essential to ensure a good performance of the WLANs. It supposes a good estimation of the number of APs, their locations according to the structure of the building, a good channel distribution and an adequate level of transmission power in order to avoid overlapping but providing the largest coverage. Otherwise, a WLAN may be composed by more access points, so it may be more expensive, but with a worse function due to the radio overlapping among APs in the same channel. In this paper, we show how a WLAN can be redesigned in order to improve its wireless coverage and function. It is based on studying the distribution and features of a public building in a Spanish University in order to determine the optimum access point location and to assign the appropriated channel. In this case, this WLAN allows users to connect to one of the available SSIDs in the target building. Results obtained from the proposed redesign have been very successful from the point of view of performance and coverage.**

Keywords- WLAN redesign; radio coverage; indoor study, WLAN; IEEE 802.11g; channel assignment.

I. INTRODUCTION

One of the most important aspects in the development and implementation of wireless networks is to ensure the optimal access to all network resources. Nowadays, ubiquitous connection to the network services is essential to let the workers and students perform their tasks. Wireless networks are widespread in both the enterprise and academia environments, providing support for wired networks and mobility to users. Wireless networks allow them to access all the services offered by the institution even when they are moving. Users can access to all resources in the infrastructure.

It is very important to know the behavior of signals within a building. An analytical analysis of the signals generated by the access points (APs) can help us to improve network coverage within the building [1]. In this new paper, we enhance our analytical study in order to propose a redesign of the wireless network of the Centre of resources for the research and learning (CRAI) of the Higher Polytechnic School of Gandia, a campus of the "Universitat Politecnica de Valencia" (UPV) in Spain. We propose a new distribution of the locations of APs with a new channel scheme. Moreover, we have included a study of wireless coverage in order to compare both values. In this comparison, we show that wireless signals in indoor environments have a different behavior, which could be used for other purposes.

One of the most important things which must be considered when a wireless network is being designed is the wireless signal losses. They depends on the number of walls and obstacles crossed in its propagation path, the materials used in the building construction, the type of obstacles, the multipath effect, and others electromagnetic waves from others systems which interference with the wireless signal. But, generally, only the fixed obstacles and walls are taken into account in the design process. There are several materials such as metal or wood very used in buildings or simply through normal walls and floors, which affect to wireless signals reducing significantly their signal level [1, 2]. In contrast, interferences caused by electromagnetic waves from other systems can be reduced selecting the most appropriate frequency band for the wireless network and a good channel assignment

When a wireless network is designed in a specific environment, it is necessary to study the distribution of the place in order to determine the better location for each AP and the channel distribution. The goal is to provide the greatest possible coverage but avoiding overlapping among channels according to the building distribution. Obviously, it is impossible to define only one model for all places and, sometimes, it is very difficult and tedious to analyze each place in detail before installing, because each one has a different distribution and with different sources of interferences. So, although it is quite easy to estimate the area of the radio coverage in a free space, it is very difficult to calculate it in indoors since the building distributions are not uniform [3]. Moreover, the irregular disposal of the objects makes the ray tracing, which mainly affects to the multipath loses, very difficult to be controlled. However, there are several features of the indoor environments which should be taken into account in order to reach a well-designed WLAN.

Moreover, an accurate design means a good sized network, that is, only APs/routers needed to cover the service area and to obtain high efficiency must be bought.

Moreover, performing a correct and optimal design of an indoor wireless networks would subsequently let the network administrators include multiple services such as positioning and tracking of people and objects [4]. So, the coverage study showed in this paper is applicable to other research fields such as wireless sensor networks [5] where the designing process presents similar inconvenient.

In addition to all the design parameters discussed above, it is essential to analyze and study which type of traffic and users the target WLAN goes to support, and how much and how many respectively. Depending on that, it will require more or less resources, bandwidth and performance. Finally, the physical distribution must be considered in order to select the APs locations since each AP needs a power supply and a point of connection to the wired network.

In short, the key issues to design and install a WLAN are to study the physical aspects of the area where the WLAN goes to be installed, to select the type of APs which fulfill the necessary requirements according to the users and traffic estimated, to perform a coverage study to assign the most appropriated location for each AP taking into account the physical features of the building (where obstacles, power supplies and points of connection to the wired network are located), and to do a good channel distribution to minimize interferences among APs.

In this paper, we use the analytical study of the building in order to know the wireless signal behavior in the CRAI building. These measures will allow us to develop new techniques for indoor network designs. In order to validate our measures, we will perform a new analysis within another scenario and compare the results between them. We will show the wireless network redesigned and how the new APs placement provides better wireless coverage.

The rest of this paper is structured as follows. Section 2 shows some related works with radio coverage and redesign of wireless networks. Section 3 presents the scenario and the tools used to perform our measurements. Section 4 explains the results of our study drawn in coverage maps. Section 5 makes a comparative study of the three analyzed radio signals in each floor. The analytical study is included in Section 6. Section 7 shows the redesign of the CRAI wireless network from the obtained measurements. This consists on relocating the wireless devices and reassigning channels according to the new distribution of devices. In order to check our proposal, we perform another analytical study in another building and show the comparison between the results in Section 8. Finally, Section 9 summarizes the conclusion and future works.

II. RELATED WORKS

Different aspects of the WLANs' coverage have been studied in several papers. There are both empirical [7] and analytical [8] studies. On the one hand, A.R. Sandeep et al. [7] suggest an indoor empirical propagation model (IEPM) in order to predict the signal strength of an indoor Wi-Fi network. It is a predictive model based on the Wall Effect Factor (Wef) and the Wall Attenuation Factor (Waf). From this model, authors can calculate the RF coverage area before installing a WLAN and so that calculating the number of access points needed. Therefore, this model means low cost and development time. On the other hand, Eisenbl et al. [8] perform an analytical study about the best location for APs and channel assignment in order to improve the performance of a WLAN. Up to now, these features have been analyzed independently, but according to this paper the greatest optimization is achieved from studying these two features simultaneously by mathematical programming. Authors propose an integrated model in order to reach a balance between both features and to optimize the indoor design of WLANs.

Expanding analyzing studies, M. Kamenetskyt et al. [9] analyze different methods for obtaining the most optimum location for the WLAN's access points. In order to evaluate the performance of these methods, they use an objective function which maximizes the coverage area and signal quality. Then different approaches to coverage planning for WLAN systems are reviewed and the most suitable for numerical evaluation are selected. From this evaluation, authors propose a new optimization scheme based on the combination of two approaches: using pruning in order to set initial locations for access points and refining these by using either neighborhood search or simulated annealing.

Then, E. Amaldi et al. [10] present a new modeling approach taking into account the effect of the IEEE802.11 access mechanism. It influences on radio coverage due to the coverage overlap between APs and its impact on the system capacity. So, they explain and discuss novel mathematical programming models based on quadratic and hyperbolic objective functions considering this. Finally, some initial results on synthetic instances are shown.

Some abovementioned authors improve its initial approach (published in [10]) because it is difficult to tackle even for small instances. So, they propose and analyze effective heuristics in [11] to tackle hyperbolic and quadratic formulations in order to maximize the overall network capacity. It is based on a combined greedy and local search algorithms turn out to provide near-optimal solutions in a reasonable amount of time.

Following with empirical papers, Kaemarungsi and Krishnamurthy [4] study the performance of the received signal strength (RSS) from IEEE 802.11b wireless network interface cards in order to improve the indoor location systems based on location fingerprints. Moreover, they point out the influence of the users' presence on the RSS, both the proximity of the human body to antenna and its orientation. These features affect the mean value and the spread of the average RSS values. So, if the position system is deployed in an environment with people, it is essential to take it into consideration while collecting RSS values for the fingerprint. In contrast, for applications that make use of sensors without human presence, this influence shouldn't be considered.

J. Lloret et al. [12, 13] show studies about an empirical coverage radio model for indoor wireless LAN design. This model has been tested on a vast number of buildings of a great extension area with over 400 wireless APs in order to get quick successful results. The objective of the model is to facilitate the design of a wireless local area network WLAN using simple calculations, because the use of statistical methods takes too much time and it is difficult to implement in most situations. The proposed analytical model is based on a derivation of the field equation of free propagation, and takes into account the structure of the building and its materials.

Sendra et al. [14] present a comparison of the IEEE 802.11a/b/g/n variants in indoor environments in order to know which the best technology is. This comparison is made in terms of the RSS indicator, the coverage area and the measurements of interferences between channels. This study only provides data from a building. So, it is difficult to extract a generalization in the wireless signals behavior.

In [15], authors propose a new WLAN design strategy called capacity based WLAN design. The method guarantee radio coverage to the target service area and provide a

specified data rate capacity to carry the traffic demand from each user in the service area. The methodology proposed determines the number of APs, frequency channels, power level and the placement of the APs that ensure the constraints as data rate density requirements, radio propagation conditions and physical limitations, related with receiver sensitivity. Authors performed several design experiments, which show the benefits of their method over the traditional coverage – based design.

Moreover, I. Broustis et al [16], suggest that when we perform large wireless sensor network deployments, it is possible to detect large amount of interference, because their small capacities. They tell us that to improve the overall capacity of the network; we can base our proposals in the intelligent frequency allocation across APs, in the load balancing of user affiliations across APs and in an adaptive power control for each AP. In their work, they search interdependencies between the three functions in order to understand when and how to apply them to the network design. The authors performed the measures following a study based on the quantification of the effects of three optimization schemes proposed in many different scenarios. From the results, we can see that applying simultaneously the three optimization schemes is not always preferable, because it can sometimes degrade the performance by up to 24% compared to using only two of the schemes.

III. SCENARIO DESCRIPTION AND USED TOOLS

The CRAI building was built in 2007. It belongs to the Higher Polytechnic School of Gandia. It is composed of 3 floors where different services for the students are offered. It contains the library, computer labs and open access classrooms. Figure 1 shows the map of this space. It is the H building of this university campus.

Now we are going to describe the scenario where the measurements have been taken from the wireless networks and the type of hardware and software used to perform our research.

A. The building

The ground floor (see Fig. 2) contains an information desk, several staff offices, a library and a large study room with a consultation area and several group study rooms. Finally, there is a multipurpose room where events and exhibitions are sometimes held.

On the first floor (see Fig. 3) we can find several computer labs, some classrooms to perform Final Degree Projects, and others group and individual study rooms.

On second floor (see Fig. 4), there are a large library with magazines, journals, books and audiovisual resources, and some computer labs and professor offices.

B. Description of UPV Wireless Network

Higher Polytechnic School of Gandia is a campus of the UPV and shares four wireless networks with the main campus, their SSIDs are EDUROAM, UPVNET2G, UPVNET and UPV-INFO. Each one of these allows university users to access to the Internet and the university resources. Their main features are:

- UPVNET: a wireless network with direct connection to all the resources of the UPV. It requires a wireless card with configured with WPA/WPA2 security.
- UPVNET2G: a direct network connection to all resources of the UPV and the Internet. It requires a wireless card configured with WPA/WPA2 security.
- EDUROAM: this wireless network is widely deployed in universities and research centers in Europe. It provides Internet access for all their members. Users only need a username and a password from their home institution. It requires a wireless card configured with WPA/WPA2 security. This network only provides Internet access.
- UPV-INFO: this wireless network works as a consultation area. It only provides all information about how the wireless network cards must be configured in the users' devices in order to connect them to some of the abovementioned networks. It uses private IP addressing and it does not allow users to access to the Internet. A second connection is needed to access to the Internet and the UPV resources. This second connection can be a Virtual Private Networking (VPN). It should only be used by very old computers that do not support WPA encryption.

In this paper we are going to analyze three of these networks (UPVNET, UPVNET2G, EDUROAM), because these are the only ones that allow users to access to the Internet.

Figure 1. Map of Polytechnic high school of Gandia.

Figure 2. Ground floor of the CRAI bulilding.

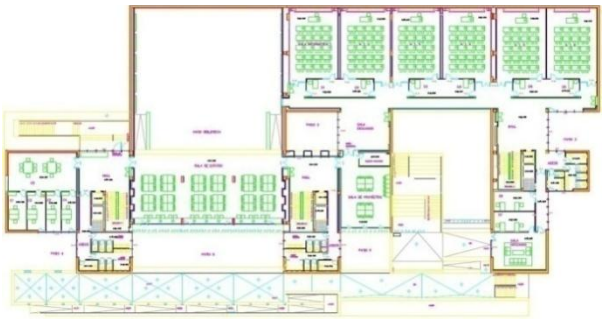


Figure 3. First floor of the CRAI building

Figure 4. Second floor of the CRAI building

C. Software and hardware used

In order to carry out this work, several measurements have been done along the three floors of the CRAI. We have used different network devices to perform these measurements:

- Linksys WUSB600N [17]: it is a USB wireless device used to gather measurements. It can capture signals from the IEEE 802.11 a/b/g/n standards. Its power transmission is 16 dBm for all standards and the receiver sensitivity is about -91dBm in both internal antennas. Transmission power consumption is less than 480mA and it consumes 300mA in the reception mode.
- Laptop: it was used to take coverage measurements. It has a dual core processor with 2 GHz per core and 2 Gbyte of RAM Memory. Its operating system is Windows Vista.
- Cisco Aironet 1130AG (AIR-AP1131AG-E-K9) [18]: this AP is the model used in all floors of the building. Its data rate can reach up to 54 Mbps. It can work at 2.4 GHz or 5 GHz, with a maximum distance from 100m to 122m in indoors (as a function of the IEEE 802.11a or IEEE 802.11g variant). The maximum distance for outdoor environments is about 198 m and 274 m. It can be powered by PoE (Power over Ethernet).

In order to capture the received signal from each selected point of the building, we used the following program:

- InSSIDer [19]: it is a free software tool which detects and controls the wireless networks and the signal strength by a graphical way. This program lists all detected wireless networks and provides several details about them such as their SSIDs, MAC addresses, channels, the radio signal strength indicator (RSSI), network type, security, speed and signal intensities which allow to control the signal qualities.

IV. COVERAGE RESULTS

We only have considered the walking area from which users typically connect to the wireless network. So, bathrooms, exterior stairways, storage rooms, etc. have been excluded. In order to perform this coverage analysis, a grid of 4 meters x 4 meters has been drawn in each floor. This allows us to take measurements for the different networks from the same places. The laptop in charge of taking measurements was located at a height of 100 cm above the ground.

A. Ground floor

This subsection shows the coverage study on the ground floor.

There are 5 APs to cover the entire plant. There are four places with the highest coverage level (the values are higher than -50 dBm). We highlight 2 rooms, Room A, the multipurpose room, and Room B, the computer room (see fig. 5, 6 and 7). The AP located outside the wall of the computer room provides coverage levels below -70 dBm inside the classroom for all three cases.

Fig. 5 shows the coverage area and levels of UPVNET wireless network on the ground floor. Room A presents signal strength of -90 dBm due to the signal attenuation suffered by the wireless signals when they cross some walls.

Fig. 6 shows the coverage area for the UPVNET2G wireless network on the ground floor. We find three places where signal strengths are higher than -50 dBm. These places are just those ones where the APs are located currently. The multipurpose room has a very low coverage on the left side because the signal is greatly attenuated by several walls.

Fig. 7 shows the value of signal strength for EDUROAM wireless network on the ground floor. Again, there are three places with signal strengths higher than -50 dBm, which correspond to the current location of the APs. In this case, more than half of the room B has signal strength levels below -70dBm.

B. First floor

This subsection shows the signal strengths measured on the first floor. In this case there are 4 APs to cover the entire plant. There are 4 places with the highest signal strengths (higher than -50dBm).

Fig. 8 shows the signal strength for UPVNET wireless network on the first floor. The rooms at the left side have low radio coverage because the AP is not located in the correct place. The offices at the right side have also very poor signal strength because they are very close to the stairs and they suffer important signal attenuation.

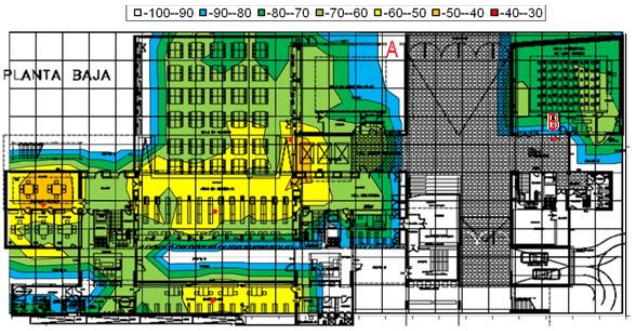


Figure 5. Radio coverage map of the ground floor for UPVNET

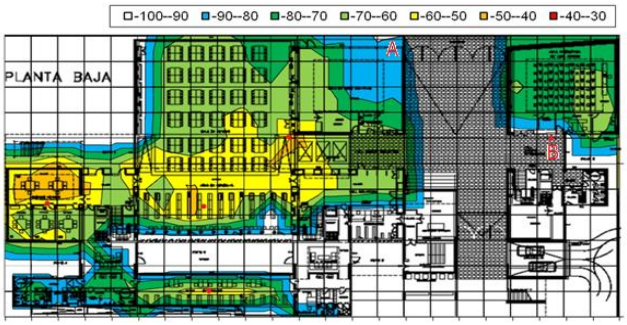


Figure 6. Radio coverage map of the ground floor for UPVNET2G

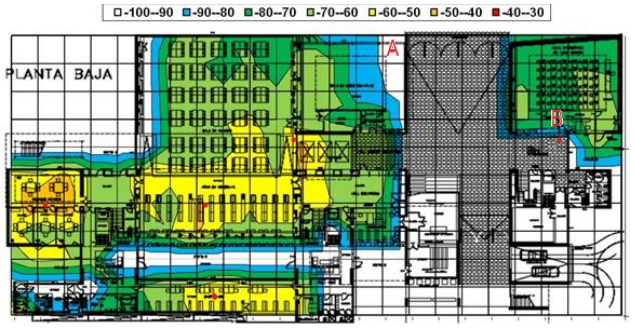


Figure 7. Radio coverage map of the ground floor for EDUROAM

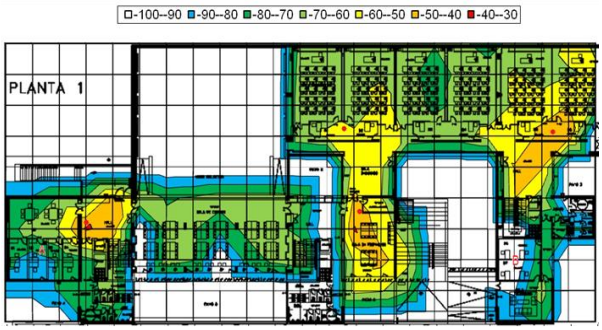


Figure 8. Radio coverage map of the first floor for UPVNET.

Fig. 9 shows the UPVNET2G wireless network signal strengths on the first floor. We can see that the classroom on the left side is not well covered because of the position of the AP. It is located on the right side of the wall. The offices from the bottom right also have very poor coverage, because they are very close to the stairs, which generate significant signal attenuation.

Fig. 10 shows the EDUROAM signal strengths on the first floor. In this case, we can see the same effect as in the other cases, but moreover there are tables in the study area (center of the picture) with low signal strength (lower than -90 dBm).

C. Second floor.

This subsection shows the signal strengths measured on the second floor. The floor is covered by 4 APs.

Fig. 11 shows the signal strengths for UPVNET wireless network on the second floor. The highest signal level is provided by the AP located at the professors offices zone (central zone of the image),which provides signal levels lower than -60 dBm. Moreover, the AP located in the hall of the two computer rooms (top right of the Fig. 11), covers virtually the entire rooms, registering levels of -70dBm in the teacher's desk. The APs located at the central-left and the bottom-left areas of the library, have signal levels around -60dBm, except at the areas near to the outer walls where values of-70dBm have been registered.

Fig. 12 shows the signal strengths from the UPVNET2G network on the second floor. In this case, the signal is propagated with levels above -60dBm, practically in both computer rooms (top right of the image). In contrast, the professor offices (central zone of the image) register levels close to -50dBm. Finally, the area of journals and audiovisual resources of the library (bottom - left of the image) presents levels around -60dBm, showing levels around to -50dBm in the area near to the AP.

Fig. 13 shows the signal strengths from EDUROAM wireless network on the second floor. The signal strength offered by the EDUROAM network is slightly lower than those ones shown for UPVNET and UPVNET2G networks. We can see that there are more areas with signal levels close to -70dBm. This happens in the computer rooms (top right of the image) and in the library (bottom - left of the image). Most of these areas are zones near to walls or walkways, which are not usually used as workplaces.

We can conclude that the signal is correctly broadcasted through the entire floor and their signal strength levels are enough acceptable to cover the working places.

After analyzing all the radio coverage images, it is easy to see that the behavior of the wireless signal in each floor is quite similar, only small variations have been registered. In addition to this, we have checked that the received signal strength is very low from bathrooms and toilets. This is because the amount of water pipes and copper tubes in the walls affects to the propagation path of the wireless signals attenuating them. We have also found low signal strength levels in the stairwells. The stairs usually are made of metal framework and a foundation which avoids a correct propagation of the signal.

Figure 9. Radio coverage map of the first floor for UPVNET2G

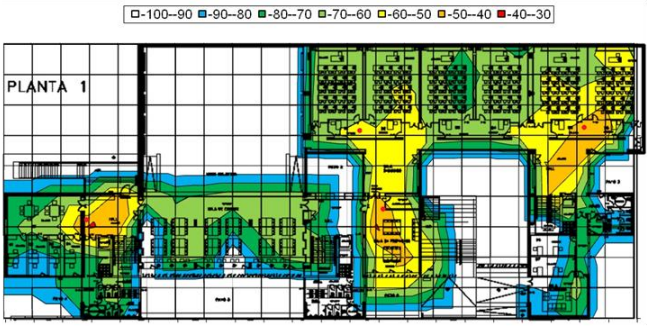


Figure 10. Radio coverage map of the first floor for EDUROAM

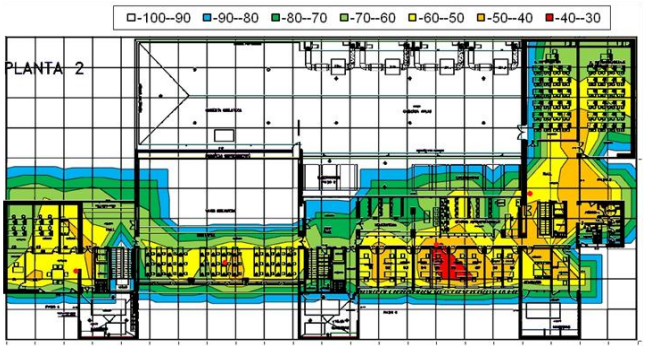


Figure 11. Radio coverage map of the second floor for UPVNET

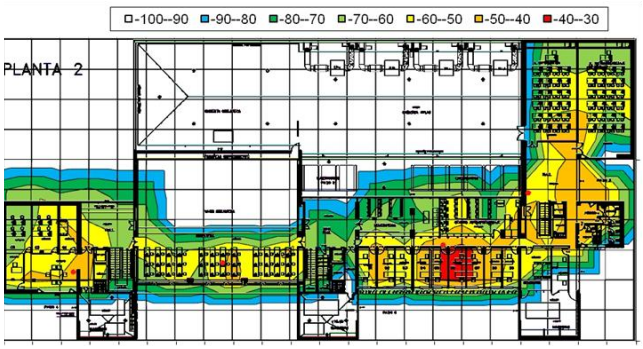


Figure 12. Radio coverage map of the second floor for UPVNET2G

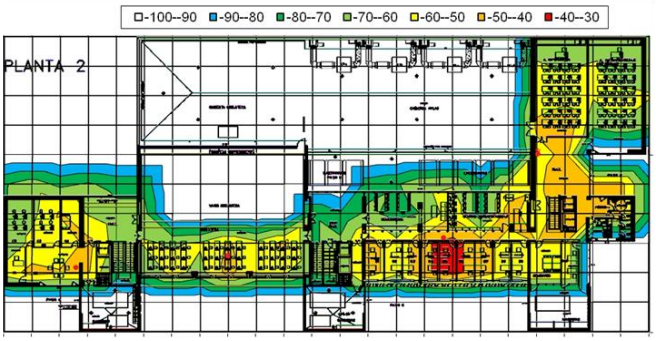


Figure 13. Signal strength on the ground floor.

V. COMPARATIVE STUDY

In this section we compare the three received wireless signals from each available wireless network on the same plant. Fig. 14 shows the three signals at the ground floor. UPVNET2G provides better signal strength levels than UPVNET and EDUROAM. Signal strengths from the first floor are shown in Fig. 15. UPVNET2G is the network which reaches the highest signal strength. UPVNET and EDUROAM show similar behaviors although there are some locations where the received signal from EDUROAM network is better.

Fig. 16 shows the behavior of signal strength on the second floor. UPVNET2G and EDUROAM show the same behavior from 3 meters to around 10 meters, but from 0 to 3 meters and from 10 meters to 12 meters, the signal strength from EDUROAM network is better. The lowest signal strength is always performed by UPVNET network. Keeping in mind all graphs, it is easy to conclude that according to signal strength, the best wireless network is UPVNET2G. Furthermore, we observe that the ground floor presents generally better signal strengths than in the other two floors.

VI. ANALYTICAL STUDY

After analyzing the above figures, we can estimate the behavior of the wireless signals in indoor environments.

Therefore, this section shows how the signal strength varies depending on the distance from the AP. In the previous section, we have shown the signal strength per floor and per SSID (Figs. 14, 15 and 16). In this section, we are going to work with the average value of all APs (per floor) and the mean value recorded for the three signals in order to analyze and generalize the overall network behavior since all APs used in the network are equal and the three signals are provided by the same AP. The mathematical equation is calculated from the tendency line of each graph.

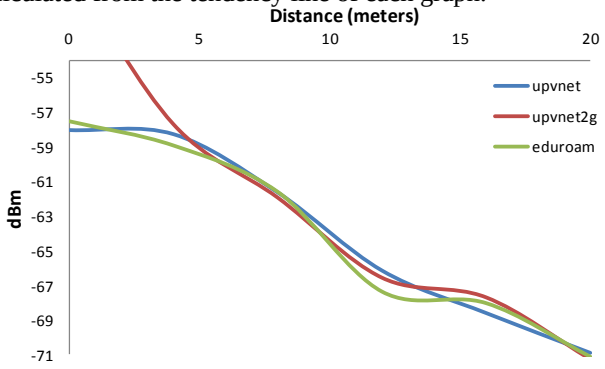


Figure 14. Signal strength on the first floor.

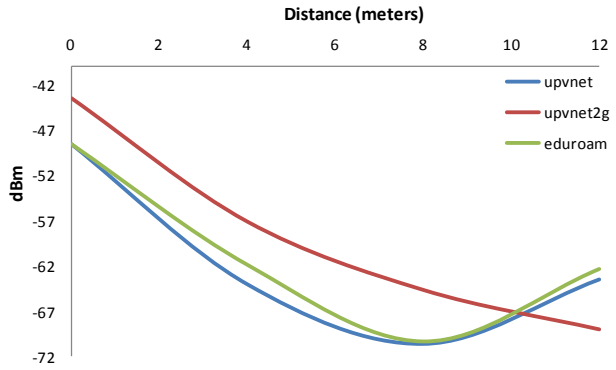


Figure 15. Signal strength on the second floor.

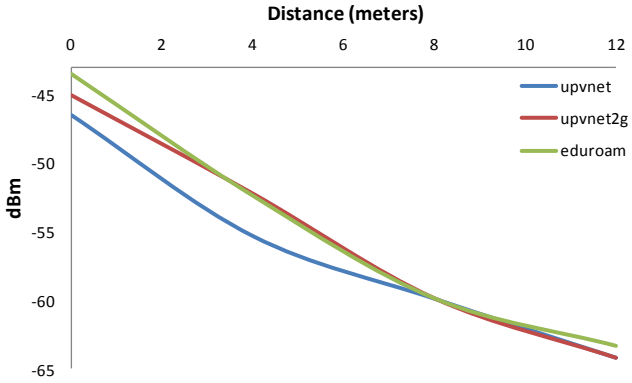


Figure 16. Radio coverage map of the second floor for EDUROAM

The analytical study is performed for three networks (UPVNET, UPVNET2G and EDUROAM) again. In order to draw each one of these graphs, we have estimated the average value of the three signals provided by each wireless network.

Fig. 17 shows the average value of the signal strength depending on the distance from the AP on the ground floor. Expression 1 shows the equation for the trend line (black line in Fig. 17) from our measurements. As we can see, it is a fifth-order polynomial equation, with a correlation coefficient (R^2) equal to 1. However, we can appreciate a slight difference between them in positions close to 3-4 meters, and further away than 17 meters from the APs.

$$Y=-0.0001x^5+0.0066x^4-0.1078x^3+0.6889x^2-2.3012x-54.75 \quad (1)$$

Where *Y* represents the average value of the received signal strength in dBm and X is the distance in meters from the AP.

Fig. 18 shows the average signal strength provided by the APs located on the first floor as a function of the distance from the APs. In positions further than 8 meters from the APs, both graphs vary very few between them, although the rest of the graph is identical. Equation 2 shows the expression for the trend line (black line in fig. 18) from our measurements.

The behavior of wireless signals based on the distance is described by a cubic polynomial with a correlation coefficient (R^2) equal to 1.

$$Y=-0.0117x^3+0.0665x^2-3.9909x-46.833 \quad (2)$$

Where Y is the signal level in dBm and X is the distance in meters from the AP.

Fig. 19 provides the behavior of the signal strength on the second floor. Equation 3 shows the trend line (black line in fig. 19) from our measurements. In this case equation 3 is a third-order polynomial equation with a correlation coefficient (R^2) equal to 1. As its correlation coefficient shows, both graphs have a nearly perfect match,.

$$Y = 0.0021x^3 + 0.0292x^2 - 2.2229x - 45 \quad (3)$$

Where Y is the average signal value in dBm and X is the distance in meters from the AP.

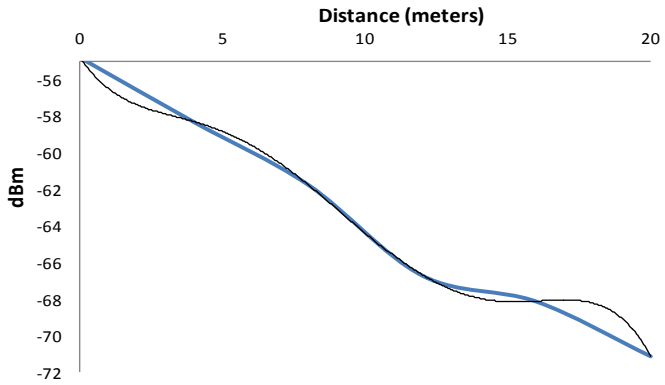


Figure 17. Average signal strength on the ground floor

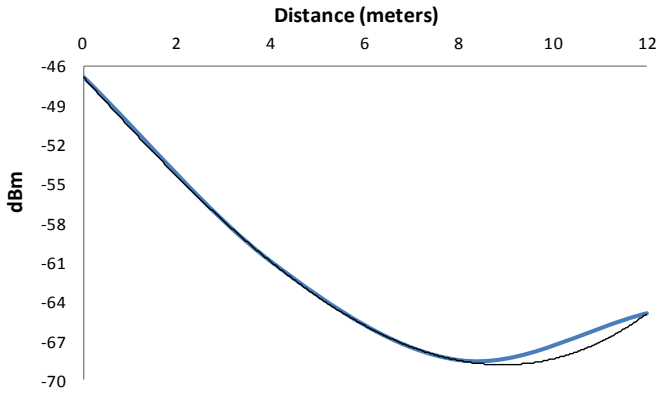


Figure 18. Average signal strength on the first floor

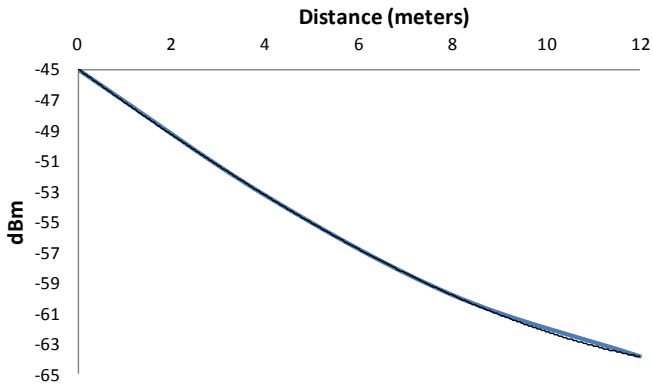


Figure 19. Average signal strength on the second floor

VII. WLAN REDESIGN

Designing WLANs which main focus is to provide the best service using the available resources efficiently requires careful planning. WLANs can be as small as a home network or as large as a network of a company with complex distributions and several buildings. Before installing a WLAN, it is essential to have the technical information about network devices and a well-defined plan for the development process. The usual sequence of steps in the APs location process is as follows:

1) Firstly, an AP is placed in each corner of each floor of the building, and it is measured the maximum coverage area for each one of them.

2) From these measurements, it is easy to determine the most suitable place for the APs taking into account that it is only needed 15% of overlapping area.

This is the most reliable process in order to find out the best location for the APs. However, it is very tedious, impractical and it implies too much time for the networks designers. Moreover, sometimes it is unfeasible for example in big buildings due to the number of measurements needed to make an accurate decision about the best potential locations or in buildings where it is not possible to gain access to take measurements [20].

We are going to relocate APs and redesign a WLAN already installed, because as we have seen in Section 4, there are some areas where wireless coverage is very poor with wireless signal strength lower than -70dbm. So, in this section we propose several changes to improve the network infrastructure and to guarantee the greatest coverage in such areas. Moreover, wireless channels used by APs are also redefined in order to reduce the interference between them.

A. WLAN Planification

In order to relocate the APs, we have taken into account the structure and distribution of the building. This is so, because there are some manufacturing materials which attenuate wireless signals significantly, for instance metal or wood. So, there are environments such as bathrooms or large fitted wardrobes where APs must not be close to.

According to the results and coverage maps shown in section 4, we decided to resign our network in order to improve its performance. The new APs placement is shown in fig. 20. Thus, several APs have been relocated on the ground floor. Firstly, the AP placed in the study room (point 1) is moved to the window and a new AP is added just in the opposite wall (point 2). This relocation has involved a better service for students who connect to the Internet in the study room because wireless coverage has been increased. Secondly, the AP located at the hall of the free access computer room is moved to beside wall in the same room. Its current position provided very low signal strength to this room. The estimated signal strengths in the ground floor are shown in fig. 21.

Moreover, on the first floor an AP (point 1) has been also relocated and two new APs have been added. The AP located in the hall on the left side (point 1) is moved to the door of the bookable study rooms. In this way, wireless signals pass through fewer walls and consequently they suffer less attenuation. The AP in point 2 is added in the center of the study room. This is a room where many students often go to work with their computers. As we have seen in the coverage maps on the first floor (section 4), some areas of the computer rooms do not have enough signal level, so we decided to add another AP in point 3 in order to improve the coverage. Fig. 22 shows the new APs location. The estimated signal strengths in the first floor are shown in Fig. 23. In both cases where APs have been relocated, we can observe that the signal strengths have been improved.

In contrast, we have not changed anything on the second place since we have checked that the signal strengths are good enough. Fig. 24 shows the position of the APs on the second floor.

Figure 20. Relocation of APs on the ground floor

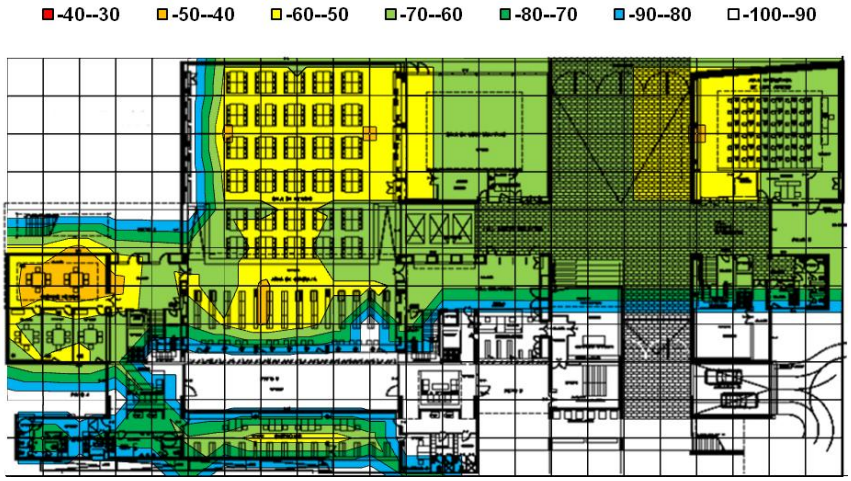


Figure 21. Estimated signal strength on the ground floor for the proposed AP relocation.

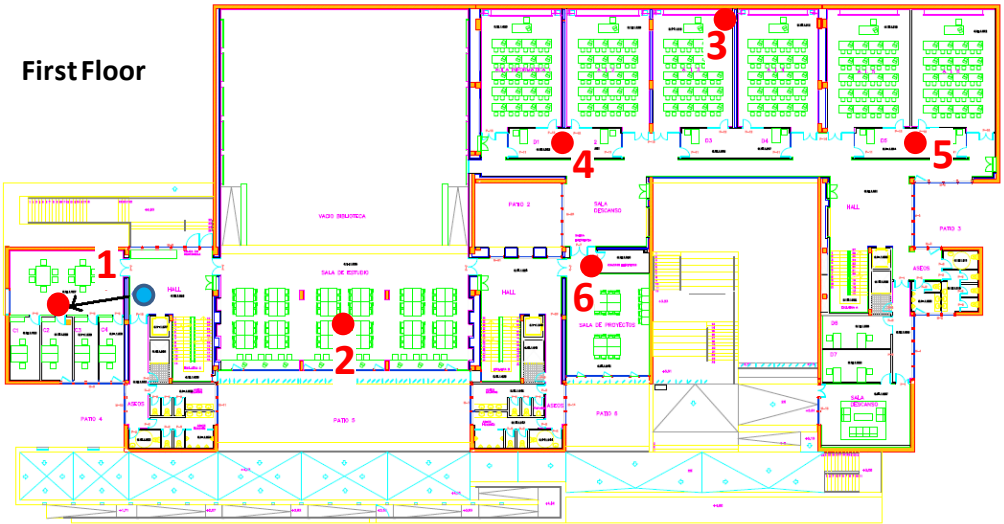


Figure 22. Relocation of APs on the first floor

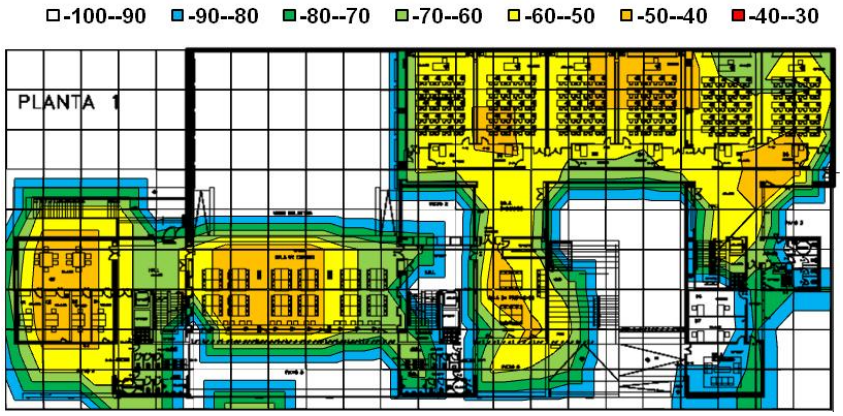


Figure 23. Estimated signal strength on the first floor for the proposed AP relocation.

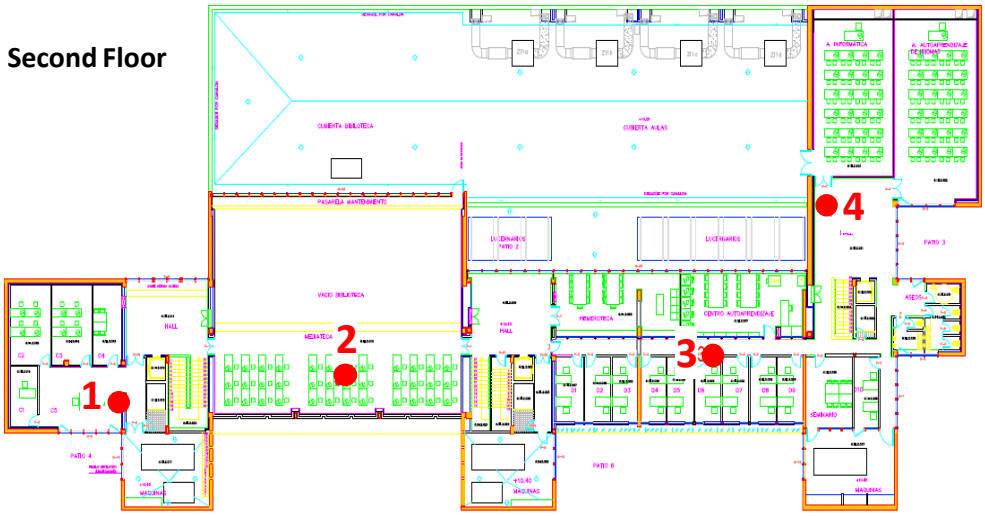


Figure 24. Relocation of APs on the second floor

B. Channel assignment

IEEE 802.11b and IEEE 802.11g standards define up to 14 channels available for wireless devices. But each country or geographic area applies their own restrictions regarding to the number of available channels. The channels are not completely independent because each channel overlaps and causes interference to the nearest four channels. Signal bandwidth (22MHz) is greater than the distance between consecutive channels (5MHz). For this reason, a gap of at least 5 channels is needed to avoid interference between adjacent cells. Using a gap of 5 channels means reaching a difference of 25MHz. Channels 1, 6 and 11 are usually the most used but the use of channels 1, 5, 9 and 13 in European domains is not bad for the performance of networks [2].

Fig. 25 shows the channel distribution from 2.412 to 2.484 GHz for the 5 main regulatory domains. In our case we use the regulatory domain of Europe, Middle East and Africa (EMEA) [13].

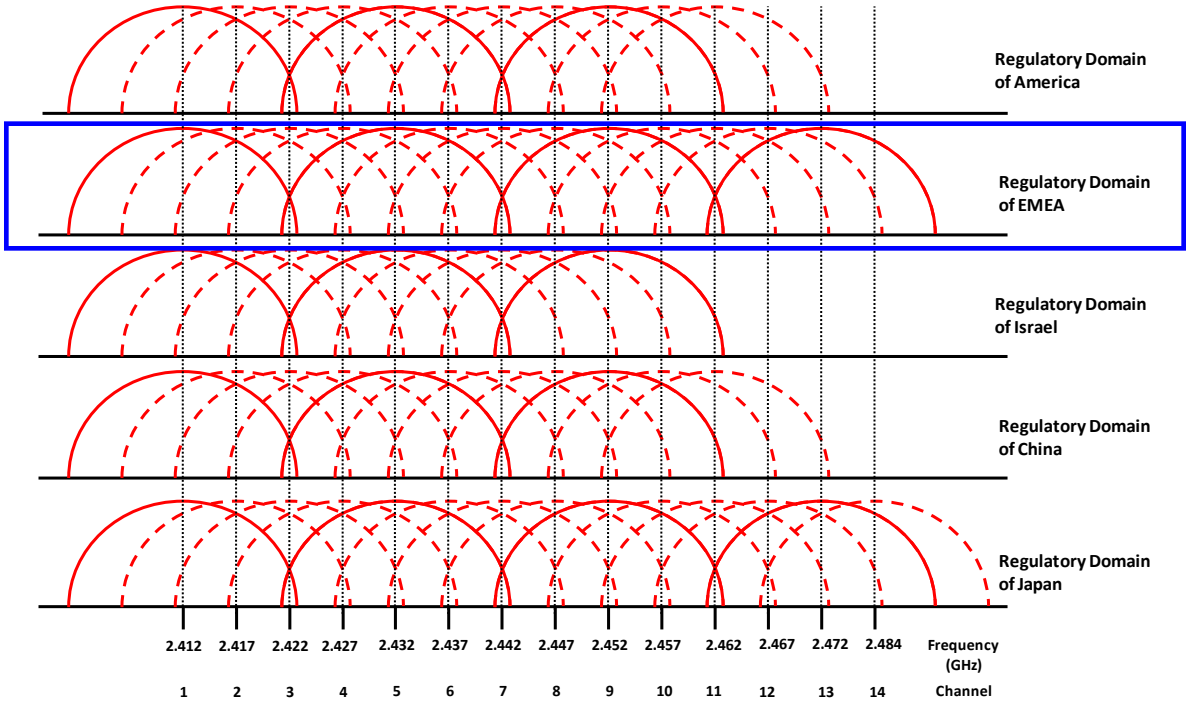


Figure 25. Channels in the frequency band of 2.4 GHz

TABLE I. CHANNEL DISTRIBUTION FOR THE GROUND FLOOR

Access Point	Channel distribution for ground floor	
	Proposed Channel	*Current Channel*
1	1	11
2	5	AP nonexistent
3	13	11
4	9	1
5	13	1
6	1	5

TABLE II. CHANNEL DISTRIBUTION FOR THE FIRST FLOOR

Access Point	Channel distribution for the ground floor	
	Proposed Channel	*Current Channel*
1	5	5
2	1	AP nonexistent
3	5	AP nonexistent
4	9	5
5	1	8
6	13	5

TABLE III. CHANNEL DISTRIBUTION FOR THE SECOND FLOOR

Access Point	Channel distribution for the ground floor	
	Proposed Channel	*Current Channel*
1	1	5
2	5	1
3	6	7
4	13	1

When new APs have been added to the network, we should define a new channel assignment for the network in order to avoid interference between the already installed APs and the new ones.

Thus, our proposal of channel assignment according to our relocation and considering one more AP for the ground floor is shown in table 1. Moreover, it can be compared with the current distribution channel in which the APs 1 and 3 share the same channel (channel 11) and the same happens with APs 4 and 5 in channel 1. This can make that these devices are interfering to each other. In our proposed channel assignment, these interferences will not occur.

In the same way, table 2 shows the channel assignment for the first floor before and after relocating the APs and adding two more. We can see that the APs 1, 4 and 6, was initially working on channel 5, while in our proposed

channel assignment the interferences between devices are insignificant.

Finally, table 3 shows the channel assignment for the second floor where no change was needed. In this case, we change the channel allocation, only to not interfere with the devices of the lower floor.

As three tables show, the current channel assignment is made so that the devices do not interfere with the devices of the same floor. However, with the proposed channel scheme in this paper, interferences between floors are also avoided.

VIII. CURRENT PERFORMANCE OF THE WIRELESS NETWORK.

In order to measure the performance of the redesigned wireless networks in this study, we monitor the performance of them during the month of November 2012, both the 2.4 GHz band and the 5GHz band. As we can see in the graphs shown in Figures 26-35, the amount of traffic carried on each band is vastly different, nearly all traffic is managed in the 2.4 GHz band, although the traffic in the 5GHz band is significant too. Moreover, we can also see how the traffic is greatly reduced during weekends.

In figure 26, we can see the evolution of the amount of input bytes in both bands and the output bytes in figure 27. Figure 28 shows the total amount of transmitted fragments by the wireless networks. Then, the figures 29 and 30 show the evolution of the amount of frames CTS (clear to send) which are received and which are not received respectively in response to an RTS (Request to Send). It should be pointed that this evolution in the 5 GHz band is nearly null. Moreover, figure 31 shows the amount of transmission's retries and figure 32 presents the amount of transmission's multiple retries. Then, figure 33 show the evolution of the number of frames with some error in the FCS (frame check sequence) and figure 34 the evolution of failed frames in general. Finally, the last one (fig. 35) show the evolution of the number of ACK which are not received when expected. We represent the evolution of these frames to show the performance of the wireless networks for each available frequency band.

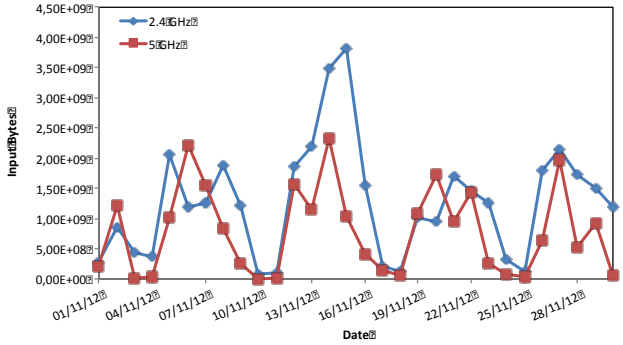


Figure 26. Input Bytes during November 2012

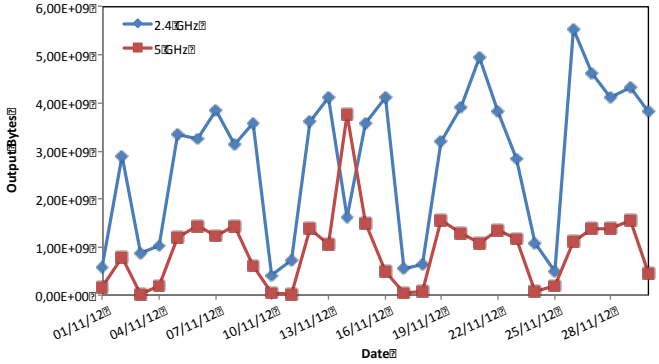


Figure 27. Output Bytes during November 2012

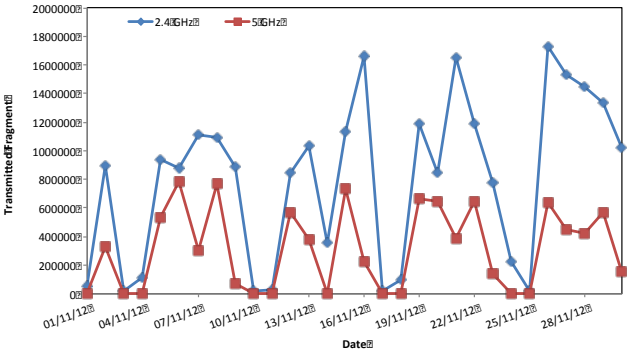


Figure 28. Transmitted Fragment during November 2012

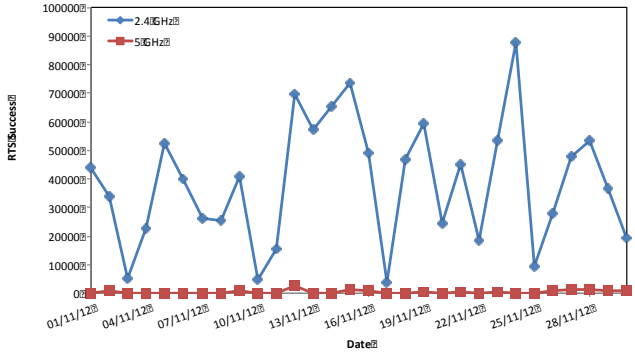


Figure 29. RTS Success during November 2012

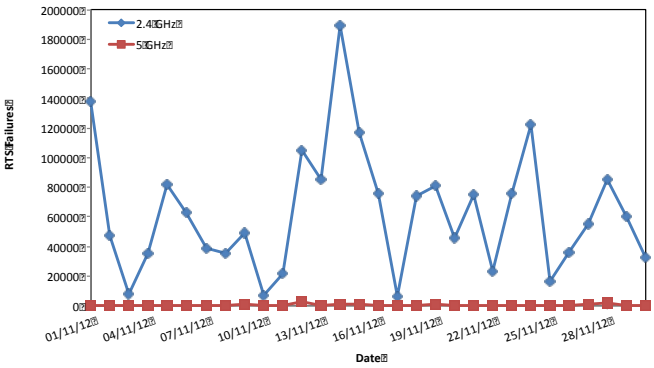


Figure 30. RTS Failures during November 2012

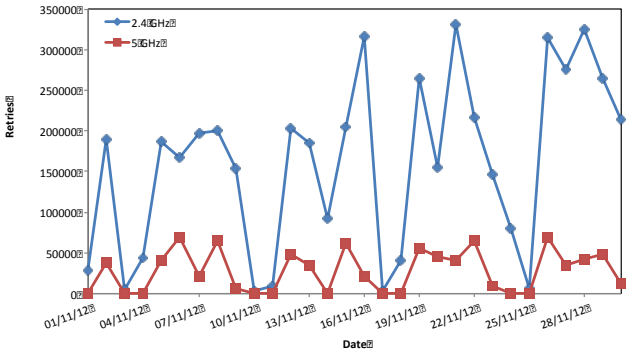


Figure 31. Retries during November 2012

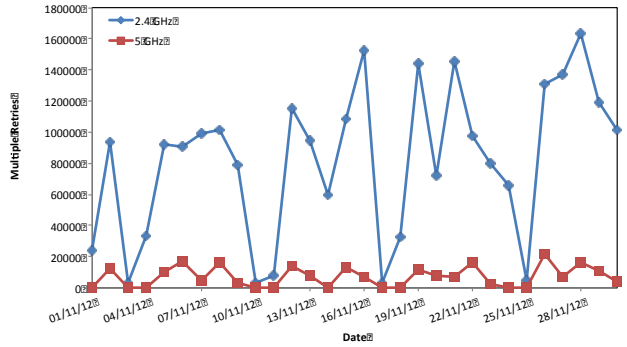


Figure 32. Multiple Retries during November 2012

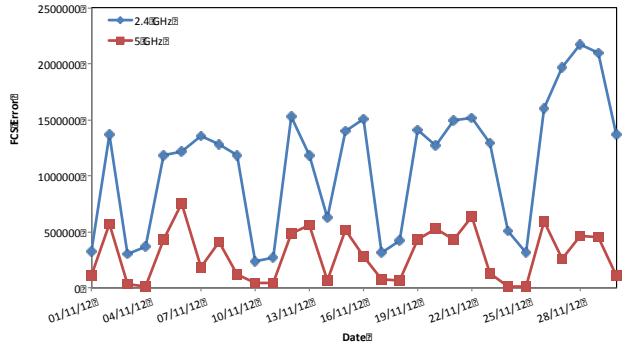


Figure 33. FCS Error during November 2012

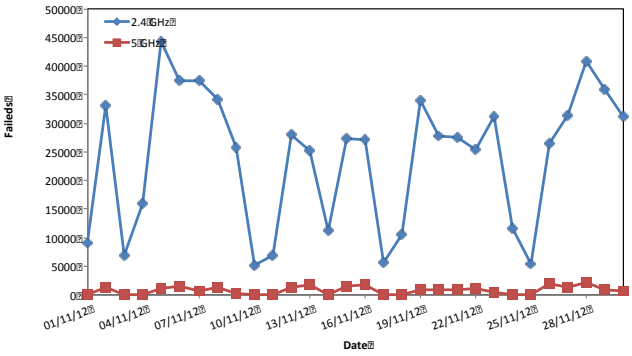


Figure 34. Failed during November 2012

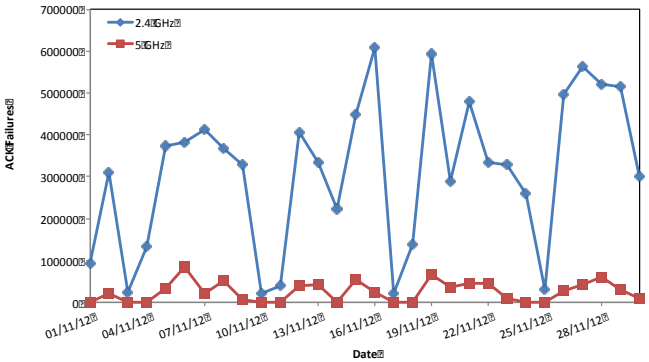


Figure 35. ACK Failures during November 2012

IX. COMPARISON WITH OTHER INDOOR COVERAGE STUDY

In order to check our study, the same process was performed in another building. In this case, only one AP was located within the building and signal strengths were measured from it. In this section, we see a comparison between both studies.

For these tests, we have considered a wide indoor environment of 91 m^2, with a length of 12.5 m., a width of 6.68 m. and a height of 2.30 m. This building is made of walls with different thickness and materials, as we can find in common houses. The plant has a rectangular base divided into two parts by a wall of 9 cm: the garage on the left side and the kitchen on the right side. The enclosure of the staircase is made of bricks with high consistency. All walls have a layer of plaster and paint on both sides. The bathroom is made of hollow bricks of 9 cm. These walls are covered by ceramic tiles. All external walls are double with a thermic and acoustic insulation of polystyrene of 5cm.

Fig. 36 shows the level of coverage obtained in IEEE 802.11g. In this case, we can see that the stairwell acts as a waveguide and signals are propagated easier in this direction than through the walls of the sides [21]. In this study, we conducted a study similar to that presented above. From that initial study, several conclusions were drawn regarding the signal behavior. However, we did not have other studies and it was difficult to generalize this behavior.

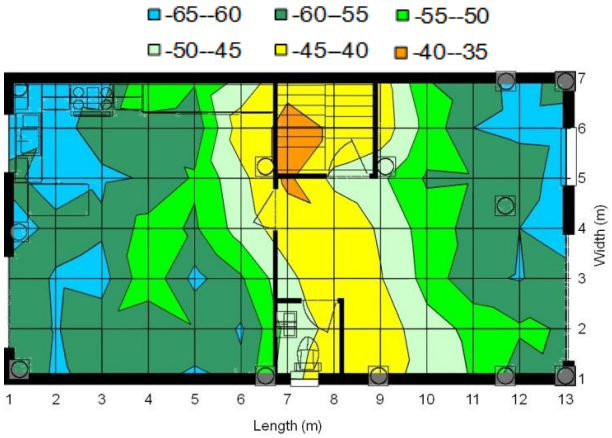


Figure 36. Coverage for the IEEE 802.11g

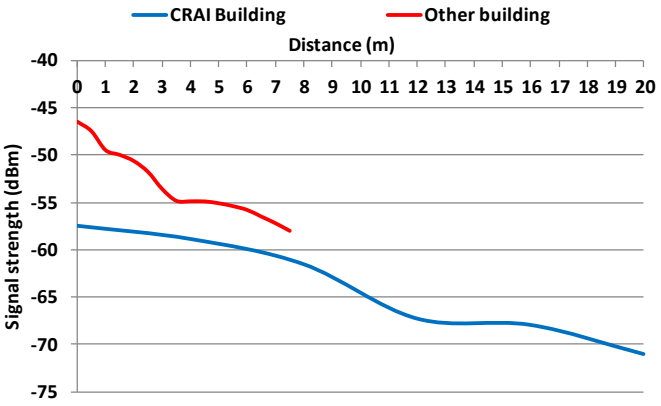


Figure 37. Signal strength for the IEEE 802.11g

For our comparison, we should observe the signal propagation through the walls, as it can be found in any building.

In this case, we can approximate the behavior of the signal by a sixth-order polynomial equation (see Eq. 4) with a correlation coefficient R^2 higher than 0.9999.

$$Y = -0{,}0356x^6 + 0{,}7695x^5 - 6{,}5329x^4 + 27{,}428x^3 - 57{,}836x^2 + 49{,}863x - 54{,}183 \quad (4)$$

where *Y* represents the average of received signal strength in dBm and X is the distance in meters from the AP.

The coverage maps shown in section 4 allowed us to analyze and characterize the behavior of the signals within this building. We performed an estimation of the average signal strength provided by the access points depending on the distance (considering the three floors). Equation 5 shows the behavior of signals when they pass through walls:

$$Y = 4 \cdot 10^{-5} \cdot x^5 - 0{,}0024x^4 + 0{,}0455x^3 - 0{,}2065x^2 - 1{,}966x - 50{,}792 \quad (5)$$

where Y represents the signal strength in dBm and X, the distance from the AP in meters.

Fig. 37 shows the behavior of the wireless signal working in IEEE 802.11g as a function of the distance. The red line shows the average of the signal strength in the new scenario and the blue line shows the average of the signal strength in the CRAI building. The most important conclusion is that the behavior of the signal is similar in both cases. However the signal strength values are different in both cases. This fact might suggest that the building area can influence in the coverage that an AP might offer. It is also shown that there are two flat zones that maintain the signal level in both signals and they could be used for future applications. We think that this difference is given by the multipath effect. In the CRAI building the signal levels (through walls) are acceptable in ranges up to 18 meters (having around -70dBm). If we extend the graph of the results obtained for the new scenario presented in this section, we estimate that we can achieve signal strengths higher than -70dBm up to 13-14 m.

X. CONCLUSION

The process to design a WLAN indoors could be a complicated and long process. It could happen that once the entire process has been performed, the result is not as successful as expected. Most probably some areas of low signal level may be created, where users do not have access to the network and to its resources.

In this work, we have performed an analytical analysis based on the signal strengths in order to enhance the performance of the WLAN of the CRAI building. We have analyzed the 3 floors inside the building and, as we have seen in section 4, some areas had low signal strength. So, from the coverage maps and using an analytical study, we redesigned the wireless network, establishing new locations for the APs and a reassignment of channels. Thus, we have improved considerably the wireless coverage for accessing to the network.

Therefore, this study can help to redesign wireless networks in similar buildings avoiding long and laborious processes. Moreover, the most suitable location for APs reduces the number of devices required to cover the whole building.

Nowadays, we are working with these measurements to propose a new indoor positioning system for wireless sensors which will allow us to monitor an environment more efficiently. Moreover we will add algorithms that mix the received signal strength indicator (RSSI) and the link quality indicator (LQI) in order to estimate the position inside the buildings [22]. We also want to expand this study for the IEEE 802.11n standard in order to reach higher data rates and greater distances.

REFERENCES

[1] S. Sendra, L. Ferrando, J. Lloret and, A. Canovas, "Indoor IEEE 802.11g Radio Coverage Study". In proceedings of the Sixth International Conference on Digital Society (ICDS 2012), pp. 121-126, Valencia (Spain), January 30 - February 4, 2012.

[2] J. Lloret, J. J. López, and G. Ramos,"Wireless LAN Deployment in Large Extension Areas: The Case of a University Campus", In proceedings of the International Conference on Communication Systems and Networks 2003, Benalmádena, Málaga (España), September 8-10, 2003.

[3] N. Pérez , C. Pabón , J.R. Uzcátegui and E. Malaver,"Nuevo modelo de propagación para redes WLAN operando en 2.4 Ghz, en ambientes interiores". TÉLÉMATIQUE 2010, Vol.9, Issue: 3, Pp.1-22.

[4] B.S. Dinesh, "Indoor Propagation Modeling at 2.4 Ghz for IEEE 802.11 Networks". Thesis Prepared for the Degree of master of science university of north texas. December, 2005.

[5] K. Kaemarungsi and P. Krishnamurthy, "Properties of Indoor Received Signal Strength for WLAN Location Fingerprinting". In proceedings of The First Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services 2004 (MOBIQUITOUS 2004), pp. 14-23, Boston, Massachusetts, USA, August 22-26, 2004.

[6] R. Mulligan, and H. M. Ammari, "Coverage in Wireless Sensor Networks: A Survey", Network Protocols and Algorithms. Vol. 2, No. 2, pp. 27-53, 2010.

[7] A. Sandeep, Y. Shreyas, S. Seth, R. Agarwal, and G. Sadashivappa. "Wireless Network Visualization and Indoor Empirical Propagation Model for a Campus WI-FI Network", World Academy of Science, Engineering and Technology, vol. 42, pp.730-734, 2008.

[8] A. Eisenblatter, H.F. Geerdes, I. Siomina, "Integrated Access Point Placement and Channel Assignment for Wireless LANs in an Indoor Office Environment", IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks (WoWMoM 2007), pp.1-10, Helsinki, Finland, June 18-21, 2007.

[9] M. Kamenetsky, and M. Unbehaun. "Coverage Planning for Outdoor Wireless LAN Systems", International Zurich Seminar on Broadband Communications, Access, Transmission, Networking, pp. 49-1 – 49-6, Zurich, Switzerland, February 19-21, 2002

[10] E. Amaldi, A. Capone, M. Cesana, and F. Malucelli, "Optimizing WLAN Radio Coverage", Procedings of the IEEE International Conference on Communications, vol. 1, pp. 180-184. June 20-24, 2004.

[11] E. Amaldi, A. Capone, M. Cesana, L. Fratta and F. Malucelli, "Algorithms for WLAN Coverage Planning", Wireless systems and mobility in next generation Internet, Springer Berlin / Heidelberg, vol. 3427/2005, pp. 52-65, 2005.

[12] J. Lloret, J. J. López, C. Turró and S. Flores, "A Fast Design Model for Indoor Radio Coverage in the 2.4 GHz Wireless LAN", in proceedings of 1st International Symposium on Wireless Communication Systems 2004 (ISWCS'04), pp. 408-412, Port Louis (Mauricio Island), September 20-22, 2004.

[13] J. Lloret and J.J. López, "Despliegue de Redes WLAN de Gran Extensión, el Caso de la Universidad Politécnica de Valencia", XVIII Simposium Nacional de la Unión Científica Internacional de Radio, A Coruña, Spain, September 10-12, 2003.

[14] S. Sendra, P. Fernandez, C. Turro and J. Lloret, "IEEE 802.11a/b/g/n Indoor Coverage and Performance Comparison", In procedings of the 6th International Conference on Wireless and Mobile Communications (ICWMC 2010), pp.185-190, Valencia, Spain, September 20-25, 2010.

[15] C.Prommak, J. Kabara, D. Tipper and, C. Charnsripinyo, "Next generation wireless LAN system design", In proceedings of Military communications conference (MILCOM 2002). Anaheim, California, October 7-10, 2002.

[16] I. Broustis, K. Papagiannaki, S.V. Krishnamurthy, M. Faloutsos, V.P. Mhatre, "Measurement-Driven Guidelines for 802.11 WLAN Design," The IEEE/ACM Transactions on Networking, vol.18, no.3, pp.722-735.

[17] Data sheet WUSB600N. Available. in: http://www.linksysbycisco.com/EU/es/products/WUSB600N, (Last Access: Aug. 2012).

[18] Specifications of cisco Aironet 1130 AG, Access Point. Avaliable in: http://www.cisco.com/en/US/prod/collateral/wireless/ps5678/ps6087/product_data_sheet0900aecd801b9058.html (Last Access: Aug. 2012).

[19] Web page of inSSIDer. Available in: http://www.metageek.net/products/inssider (Last Access: Aug. 2012).

[20] W. Stallings, "Data and Computer Communications", (7th ed.), Prentice Hall PTR, 2004.

[21] S. Sendra, M. Garcia, C. Turro and, J. Lloret, "WLAN IEEE 802.11a/b/g/n Indoor Coverage and Interference Performance Study", International Journal on Advances in Networks and Services, vol. 4, no. 1 & 2, pp.209-222, 2011.

[22] Talmai Oliveira, Madhanmohan Raju, Dharma P Agrawal, Accurate Distance Estimation Using Fuzzy based combined RSSI/LQI Values in an Indoor Scenario: Experimental Verification, Network Protocols and Algoprithms, Vol 4, No 4 (2012), Pp. 174-199.

Enabling User Involvement in Trust Decision Making for Inter-Enterprise Collaborations

Puneet Kaur, Sini Ruohomaa, and Lea Kutvonen
Department of Computer Science
University of Helsinki
PO Box 68, FI-00014 University of Helsinki, Finland
puneet.kaur@cs.helsinki.fi, sini.ruohomaa@cs.helsinki.fi, lea.kutvonen@cs.helsinki.fi

***Abstract*—Trust decisions on inter-enterprise collaborations involve a trustor's subjective evaluation of its willingness to participate in a specific collaboration, given the risks and incentives involved. We have built support for automating routine trust decisions based on a combination of risk, reputation and incentive information. To handle non-routine decisions, we must provide human users with a way to interface with this system and gain access to supporting information. Current collaboration management systems are missing the concepts, processes and interfaces for enabling user involvement. In this paper, we present two key contributions towards enabling user involvement in trust decision making for inter-enterprise collaborations: i) We have studied existing literature on human trust decision making perspectives, and produced a set of criteria for trust decisions. We analyze how three collaboration management systems support these criteria. ii) We provide a more detailed case study of enabling these features in our Pilarcos collaboration management system through implementing a trust decision expert tool prototype, and report the results of our user experiments on it.**

Keywords-trust decisions; inter-enterprise collaborations; collaboration management middleware; user interfaces

I. Introduction

Networked business is moving from closed strategic networks into open business ecosystems where inter-enterprise collaborations, i.e., business networks, are facilitated. In an inter-enterprise collaboration, services from independent organizations are brought together by interrelated business processes to achieve a shared goal for end users as a composed service. An online travel agency, for example, can compose travel packages for its customers by utilizing a set of services provided by its partners for payment handling, booking flights, hotel itinerary, car rental and other location-specific arrangements, where each partial service is provided by a separate autonomous enterprise [1].

We define a service as a network-accessible object with a published interface. It can consist entirely of software, or be a software interface to a physical system, such as a logistics service that can be reserved online. Inter-enterprise collaborations composed of these services involve multiple interdependent parties, extending beyond isolated provider and consumer interactions.

Inter-enterprise collaborations are particularly useful for small and medium-sized enterprises, which hold expertise in their own domain but have limited resources. By collaborating with other enterprises, they can attain a competitive edge in fields outside their individual scope, and also join forces to expand their business into fields dominated by large enterprises [2], [3]. Large enterprises can apply the same methods to organize their production life-cycles in-house, or to experiment on new service concepts together with external collaborators.

As the demand for easily set up collaborations between interoperable services grows in both the number and the scale of the collaborations themselves, ad hoc collaboration establishment and decision-making solutions are no longer cost effective.

The success of inter-enterprise collaborations relies on dynamically evolving open service ecosystems, supported by a flexible infrastructure that reduces the cost of setting up and managing the collaborations. This infrastructure ensures that individual enterprises do not need to solve issues of interoperability management, collaboration coordination, breach recovery and trust management using costly manual administration solutions.

The emergence of technology support, ranging from service-oriented architecture and Web Services to cloud infrastructures, are paving the way for semi-automated and low-cost setup and management of inter-enterprise collaborations. The Pilarcos open service ecosystem that we have proposed in earlier work [3] provides infrastructure services for finding potential partners and ensuring service interoperability, collaboration management and semi-automated trust decisions [2], for example; in this paper, we focus on the trust management support specifically.

The service provider enterprises operating in open service ecosystems are autonomous, and new service providers can join the ecosystem to offer new types of services by publishing service offers. The continuously evolving group of service providers and their independence makes trust management both important and challenging. Fine-grained routine decision-making and the monitoring to support the decisions should be automated to not form a bottleneck

in otherwise highly automated collaboration management, but the organizational and individual users' decision-making process and the conceptual basis required by them are not sufficiently supported by automation tools yet.

Automated trust decisions can only be relied on in routine cases: human intervention is required for making trust decisions in situations where the risk or incentives are particularly high, or the information available for supporting the decisions is insufficient. Current collaboration management systems are missing the concepts, processes and interfaces for enabling user involvement.

In earlier work, we have set the basis for how to identify points where human intervention is needed through metapolicy, including support for measuring the amount and quality of the input information [2], [4]. We have also analyzed different existing models for the trust building of individual [5] and organizational users.

In this paper, we present two key contributions towards enabling user involvement in trust decision making for inter-enterprise collaborations: i) We have studied existing literature on human trust decision making perspectives, and produced a set of criteria for trust decisions. We analyze how three collaboration management systems support these criteria. ii) We provide a more detailed case study of enabling these features in our Pilarcos collaboration management system through implementing a prototype of an expert tool for trust decisions, and report the results of our user experiments on it.

The rest of the paper is organized as follows: Section II provides an overview of the problem environment of making trust decisions on inter-enterprise collaborations. Section III presents our model of human decision-making criteria, based on existing literature on human perspectives on trust, and compares it to existing collaboration management systems to analyze how well the concepts are supported there. Section IV provides a detailed case study of enabling these features in Pilarcos: we have implemented an expert tool user interface prototype and present the results of user evaluations. Section V concludes the paper.

II. Trust in Inter-Enterprise Collaborations

Inter-enterprise collaboration depends on trust, as the autonomy of partners causes uncertainty and risk, which must be found acceptable for the collaboration to be established and for it to proceed. In this section, we discuss the basis of trust decisions on inter-enterprise collaborations and present three example trust management systems, which we will compare against our human trust decision-making criteria in Section III.

A. Trust Decisions on Inter-Enterprise Collaborations

For our context of inter-enterprise collaborations, we define trust as "the extent to which one party is willing to participate in a given action with a given partner in a given situation, considering the risks and incentives involved" [2]. Trust is crucial for the sustainability, existence and stability of any inter-enterprise collaboration, particularly when possibilities of direct interaction are limited [6], [7]. The stronger the willingness to depend on and cooperate with the other party is, the less need there is for explicit risk reduction, monitoring and other protective structures. In this sense, a strong trust relationship improves the performance and overall efficiency of the established inter-enterprise collaboration. On the other hand, the open service ecosystem will inevitably have untrustworthy actors as well, and each service provider must protect themselves against such risks. The goal of trust management is to identify the appropriate level of caution for different situations.

A trust decision compares the risks and incentives involved in a given decision context; at the basic level, the outcome is either positive (yes, collaborate) or negative (no, withdraw). A reduction in perceived risk or increase in incentives can be introduced to change the result through application of additional protection mechanisms or contract negotiations, for example.

Trust decisions are made during the establishment of inter-enterprise collaborations, and routinely during their operation, whenever new resources are committed. In the establishment phase, explicit trust decisions are needed because some of the actors in the ecosystem are previously unknown or little known: we argue that not fixing the set of possible collaborators beforehand to a well-weathered strategic network of enterprises provides competitive advantage [2]. The decision made in the establishment phase does not have complete information available on how the collaboration will span out. Instead, during the operation of a collaboration, committing more resources or observing significant behavioural deviations trigger the need for new trust decisions.

Trust decisions measure the subjective willingness of a trustor to perform a given action with a given trustee. In the context of our work, both the trustors and trustees are business services; this level of abstraction reflects the fact that two services even within the same enterprise may hold different information, have a different effect on assets, and be governed by different policies. When a trust decision is delegated to a human user, therefore, the interventions are made on behalf of a specific service, not the entire enterprise.

Some of the expected risks and gains can be estimated based on the trustee's past behaviour, represented by their reputation. The balancing incentives are created by the business importance of the activity itself, such as a need to fulfill existing contracts, or a desire to try out a new way of making business or a new set of partners.

A large body of existing work on reputation systems has focused on electronic markets, aiming to support either human or automated decisions specifically. Work in this environment involves one-on-one transactions to purchase

goods or services, and often focuses on a relatively narrow view of reputation information only [8].

A common approach focusing on the reputation management aspect in particular applies Bayesian modelling to estimating the trustee's likely future behaviour. In a nutshell, this branch of research focuses effort on mathematically credible calculation of the probability of different possible outcomes for the next event based on past information. The Beta reputation system [9] is a well-known example of this group. As its name implies, the system fits binary positive or negative experiences into a beta distribution [9]. Other work building on this approach proposes ways to extend this approach either from binary to discrete scales [10], [11] or to collecting binary data on more than one dimension, such as prompt delivery and overall approval of the product [12]. Aspects such as how to weigh experiences reported by other parties by their credibility are considered as well [11], [12].

From our point of view, research on Bayesian reputation and trust management systems focuses on a mathematically elegant transformation of reputation information into a risk evaluation. In other words, Bayesian trust and reputation systems seek to find a good pair of reputation update and trust decision policies, given a relatively simple baseline information model. We find that Bayesian policies can operate within our information model [13, ch. 4.3], although they only use a part of the provided input. Within this subfield, the search for well-performing policies has led to a need to fix the problem field reasonably early, and the information model assumptions warranted by this comparability requirement may prove to be too rigid for more general use [8]. On the other hand, for automation purposes the line must always be drawn somewhere. While we have chosen a tradeoff of increased information model complexity in exchange for capturing more aspects of trust decisions that we consider important, our model remains a simplification as well.

For a broader overview on related work on trust modelling, we refer to surveys on reputation systems [14], [15] and their robustness [8], [16]. Robustness in the face of attacks and other misbehaviour is one of the key issues that demand a holistic and high-complexity model of the operational environment to properly address. As trust management falls under the category of protective systems, robustness is an essential requirement; we discuss it more extensively in earlier work [13].

We have found that inter-enterprise collaborations, which involve multiple partners and a wide range of interdependent services, require a broader information model in order to capture the variety of risks and incentives as well as their dependence on the decision context [2]. While additional factors make the system more complex to understand, they vastly improve its configurability through policy and metapolicy [13].

In addition to Pilarcos, explicit risk information and the separation of risk calculation policy from reputation management has been proposed by the SECURE project [17] for ubiquitous computing either between private users or in client-to-business settings. As SECURE aimed to produce a personal general-purpose trust decision assistant to carry with the user, it also had to address the requirement for a flexible and configurable trust information model. For Bayesian systems, on the other hand, the chosen approach seems to reflect the target application area as well: in eBay [18], a typical electronic market where goods are bought and sold, a user's most visible reputation has long consisted of counters of the positive, neutral and negative feedback they have received [19].

In a recent paper, Marsh et al. specify a set of requirements for trust models from a usability perspective [20]: understandability by the actual users, support for monitoring and user intervention, actively prompting for input in uncertain situations, extensive configurability by the user delegating the decisions to the system, catering for different time frames for when the decision responses are needed, accepting the incompleteness of available information, and allowing the user to find out more about the decision context. This call for usability of trust models matches our goals well; we find that being able to understand and control the system is a requirement for users to trust the system to handle their trust decisions for them.

In terms of state of the art in information representation for nontrivial trust models, Ries has proposed and conducted user experiments on a two-dimensional graphical representation of positive outcome probability and the amount of available information in his thesis [21]. Research shows that the explicitly presented reputation information is not the only element affecting a trust decision: users may be influenced by unexpected user interface elements as well, such as decorative images [22], [23].

B. Trust Support in Inter-Enterprise Collaboration Management Systems

In preparation for comparison of the human and organizational decision-making perspectives in various trust management systems, we briefly introduce three inter-enterprise collaboration management systems that have suitably similar goals: i) TrustCoM, ii) ECOLEAD and iii) Pilarcos. The last of these will be further detailed to provide insight for the integration of the proposed interface.

TrustCoM is a large European Union project in the domain of inter-enterprise collaborations. The main contribution of TrustCoM has been the architectural and conceptual framework addressing trust, security and contractual issues from the perspective of inter-enterprise collaborations [24], [25]. The TrustCoM framework [24], [26] supports trust decisions during the joining and continuation of the collaboration. In TrustCoM, trust decisions are made when a new partner needs to be added or a previous partner

needs to be replaced. The trust decisions are made based on reputation information measuring trustee capabilities, integrity and benevolence, in addition to functional definitions of the role, requirements of quality of service, cost and security. The TrustCoM framework also involves an initial user interface in the form of an eLearning portal in a scenario demonstrator [26], helping users find the service best suited for them. In the general case, the design of trustworthy and secure user interfaces falls outside the scope of TrustCoM, although their importance is acknowledged.

ECOLEAD is also a European Union Integrated Project. It introduces a plug-and-play, pay-per-use, platform-independent and secure ICT infrastructure for the initiation and functioning of the inter-enterprise collaborations [6], [27]. The project addresses interoperability, trust, security, transparency, and affordability [27]. In ECOLEAD [6], [27], trust decisions are made at two points: base trust is established during the entry into the ecosystem, and specific trust is evaluated when each inter-enterprise collaboration is set up. For base trust, all enterprises entering the ecosystem answer a questionnaire on, e.g., organizational competences, prior successful collaborations, prior engagement in opportunistic behaviour, and adherence to technology standards and delivery dates [28]. Collaboration-specific trust is established in a hierarchical manner, starting from the specification of objectives in terms of measurable elements. The ICT infrastructure of ECOLEAD also provides support for portlets, pluggable user interface elements, for interaction with the users [27]. The trust prototype has a web and mobile portlet, providing a list of potential partners for collaboration, where the users can select those found most suitable for the task. Like TrustCoM, ECOLEAD does not focus on user interfaces for trust decision making specifically.

The Pilarcos open service ecosystem we have proposed in earlier work [3] aims to help service providers to find and collaborate with partners from the open service market. Collaborations are defined by a chosen business network model, which defines roles consisting of specific service types that the participanting services implement, and the shared business processes. Due to some of the potential partners being unknown or little known, trust management requires explicit support.

Pilarcos includes infrastructure services for finding and selecting potential partners and ensuring service interoperability, eContracting and collaboration management, local monitoring to collect evidence of trustworthiness, and semi-automated trust decisions [2]. The distributed infrastructure services in Pilarcos allow the enterprises to make local, private trust decisions on whether they want to join or continue in an inter-enterprise collaboration. Other decision policies are also needed to ensure that the automation tools act in accordance to the strategy and privacy requirements of the enterprise, for example [2].

While TrustCoM makes trust decisions when partners are added or changed in an operating collaboration, Pilarcos trust decisions are made both at the start of a collaboration and when new resources are committed, in which case the decision may trigger partner changes as well. The trust decisions in Pilarcos compare the subjective risks, calculated on the basis of the past behaviour of the trustee as encoded in its computational reputation, and incentives involved in the endeavour, such as the business importance of the collaboration [29]. The incentives are reflected on how much risk is tolerated. Routine decisions are automated, following local policies [2]; however, there are always situations that require human intervention. To allow for this, the decision policies define risk tolerance ranges for automatic acceptance, automatic rejection, and for requesting input from the human user [1]. Possible reasons for requiring human intervention include high risks combined with high incentives, or too little reputation information available on the trustee.

III. Decision-Making Perspectives

Designing a trust decision expert tool meant for humans requires understanding the process of human trust decision making in general. The reviewed research on human needs for decision-making has been conducted both in the context of electronic commerce and more general settings. The domain of inter-enterprise collaborations specifically has remained relatively unresearched, while the focus has been on business-to-consumer (B2C) settings. Considering the underlying problem, we believe that findings regarding human trust decision making in the B2C electronic commerce domain, for example, can be applied in the case of inter-enterprise collaborations with adaptations.

In this section, we present our model of human decision-making criteria, which is based on existing literature on human perspectives on trust. We then compare the three inter-enterprise collaboration management systems presented in the previous section in order to study how well the concepts we have extracted are supported in them.

We have studied the research literature from two perspectives: First, approaches to human trust development, and second, qualitative and quantitative criteria for trust decision making.

A. Human Preferences on Trust Decision Making

1) Approaches to Human Trust Development: The existing literature reveals two different approaches to modelling human trust development: cyclic and staged [30]. Table I summarizes the reviewed approaches.

The cyclic approach to trust development was introduced by Fung et al. [31]. It relies on the development of the trustor's confidence based on the satisfaction of prior behavioural outcome expectations. However, continuous distrust at any phase has a negative effect on the existing trust levels. Fung et al. [31] and Deelman et al. [32] have proposed two different models for cyclic trust development.

Table I A SUMMARY OF THE REVIEWED APPROACHES TO HUMAN TRUST DEVELOPMENT.

Author	Approach	Summary
Fung et al.	Cyclic	• Factors affecting initial trust establishment: information quality, interface design and reputation • Future trust is based on the satisfaction of prior expectations
Deelman et al.	Cyclic	• Factors affecting initial trust establishment: willingness to trust, estimation of trustworthiness of trustee, evaluation of past experiences, situation and risk inherent in current situation • Future trust is based on the satisfaction of prior expectations
Shapiro et al.	Three-stage	• Deterrence-based trust: relies on measures preventing occurrence of misbehaviour • Knowledge-based trust: relies on a knowledge gained as a result of direct interaction with trustee, and satisfaction of prior expectations • Identification-based trust: highest trust level, relies on the outcomes or experiences gained as a result of repeated interactions with trustee
Ba	Three-stage	• Calculus trust: relies on comparison of gains versus possible losses • Information-based trust: relies on information gained as a result of direct interaction with trustee and satisfaction of prior expectations • Transference-based trust: highest trust level, relies on the outcomes or experiences gained as a result of repeated interactions with trustee
Kim et al.	Two-stage	• Initial stage: marked by either no or low trust • Commitment stage: high-trust stage, where the trust relies on prior direct interactions with trustee
McKnight et al.	Two-stage	• Exploratory stage: marked by either no or low trust • Commitment stage: high-trust stage, where the trust relies on prior direct interactions with trustee

Fung et al. [31] present the basic model where information quality, interface design and reputation are the factors contributing to the initial trust development. Building on this, Deelman et al. [32] have further elaborated the original model by proposing additional factors: willingness to trust, estimation of the trustworthiness of the trustee, evaluation of past experiences, situation, and risk inherent in the current situation. The most notable point about the model of Deelman et al. is that it is applied to trust development in the domain of inter-enterprise e-commerce.

When we consider inter-enterprise collaborations in particular, the needs of trust modelling are slightly different. As regards to domain-specific needs, we would like to enhance the cyclic models proposed by Deelman et al. [32] and Fung et al. [31] in the following three ways: First, shared vision, contracts and legal conditions also play a significant role, especially during the initial stages of trust development; they should therefore also be considered as factors affecting trust development. Second, distrust development should be covered by the models as well, as services may change their behaviour and trust relationships may degrade or be broken off as a result. Third, while Deelman et al.'s model suggests that the given factors should be followed in sequential order, we find that the factors collectively affect trust development and their order completely depends on human preference.

Staged trust development models exist in different forms, with a different number of stages proposed for trust development. Shapiro et al. [33] and Ba [34] have both proposed a three-stage trust development model. The stages proposed by Shapiro et al. are deterrence-based, knowledge-based and identification-based trust. Ba instead proposes that the stages are calculus-based, information-based and transference-based. Kim et al. [35] and McKnight et al. [36] have suggested two-stage trust development models, which both call the final stage of trust development the commitment stage. Kim et al. refer to the first stage as initial trust development, while McKnight et al. call it the exploratory stage.

The staged trust development models cover different angles, but problems remain for applying them to the domain of inter-enterprise collaborations. The three-stage models do not consider the effect of opportunistic behaviour, such as degrading quality of service or contract violations due to changes in the priorities of an enterprise. Similarly, McKnight et al.'s model does not discuss the possibility of degrading or withdrawing from the existing trust relationship. As the financial situation and motivation of an enterprise can change at any time, its behaviour can notably change as well, and we believe that the trust development models should be able to capture this.

The three-stage trust development models are also limited in the sense of assuming that trust development proceeds in a sequential order. We noted a similar issue with the sequential order in Deelman et al.'s cyclic model; in this case we find that two stages of trust development could well be in use simultaneously. For example, Shapiro et al.'s knowledge-based and Ba's information-based trust are fixed as the second stage, which occurs after a series of direct interactions. We find that information from third-party reputation networks can be used also during the first stage,

before first-hand interactions have taken place. Apart from this, the model of Kim et al. does not propose a precise list of factors affecting trust development in the initial stage, and do not discuss the shift from initial to committed stage.

2) Criteria for Trust Decision Making: The second target of our literature research involves identifying different factors that affect trust decision making. The term 'criteria' refers to the various qualitative and quantitative factors that play a significant role in the process of human trust decision making [37], [38], [30]. Criteria for trust decision making in business-to-consumer (B2C) e-commerce can relate to institutional, environmental, website, trustor and trustee metrics, for example. The different criteria are applied in trust decision making to help analyze the situation that has called for a decision. They should be considered because they are necessary to complete the human trust decision making process.

The following section presents the different criteria we have selected for trust decision making in the domain of inter-enterprise collaborations within four categories: trustor, trustee, contextual and collaboration-specific criteria. The criteria have been gathered based on the process of human trust decision making, extended to suit the needs of inter-enterprise collaborations in particular.

Trustor Criteria: The trustor criteria are attributes of the trusting entity that affect the process of trust decision making. The personality and thinking of the trustor have a major influence on the final decision. The trustor criteria are propensity to trust, emotions and culture.

Propensity to trust is defined as a human behavioural trait, reflecting their inherent willingness or attitude towards trusting humanity, independent of any information regarding the entity to be trusted [35]. It makes the trustor risk-seeking, risk-averse or risk-neutral. The establishment of propensity to trust is based on prior experiences, starting from infancy. It can also be viewed as dispositional in nature, as human beings are taking their propensity to trust from one situation to another [38]. Propensity to trust plays a significant role during the initial stages of trust establishment in the case of previously unknown or little known entities [39].

Emotions can be defined as a cognitive approach to trust decision making [40], [41]. They are also independent of any kind of information regarding the target entity. Emotions bring in "temporal irrationality" in the process of trust decision making [41]. They might instigate a positive trust decision if the person is in a happy mood even in a situation that does not warrant trust at all otherwise, for example. Emotions acquire a dominant role in trust decision making by formulating a viewpoint regarding the available information and the current situation requiring the trust decision.

Culture is another personality trait of the trustor. It impacts the trustor's attitude, which plays a significant role in perceiving available information [41]. This way it also has an influence on the other two trustor criteria. For example, a small or medium-sized enterprise may be less willing to take a certain measure of risk as compared to large enterprises, owing to limited resources affecting their general tendency to trust. Shoorman et al. [41] state that the relationship versus task dimension of culture plays a major role in the process of trust decision making. This means that a task-oriented culture will be more risk-seeking, while a relationship-oriented culture invests particularly in building and maintaining long-lasting trust relationships.

We believe that all the aforementioned trustor criteria are a good fit with the domain of inter-enterprise collaborations. Their role becomes evident when we consider human trust decision making. However, they are also reflected in the automated trust decision making through the use of private, local or mutually decided and negotiated policies, contracts, rules and regulations. The mutual decision and negotiation is being carried out through either machine agents that are administered and configured by human users, or through the human users themselves.

Trustee Criteria: The trustee is the entity targeted by trust. As trust balances for risks that are inherent and not easily eliminated from the decision-making situation, the attributes of the trustee have a major role in the process of trust decision making as well. We consider trustee reputation to be the key trustee criterion.

Reputation information refers to the knowledge about the past and present behaviour of the trustee [6], [2], [37]. When coupled with an assumption of behavioural consistency, this historical information is used for making predictions about the trustee's future behaviour, and assessing the overall trustworthiness of the trustee. According to Mayer et al. [38], trustworthiness can be perceived in terms of ability, benevolence and integrity. Here, ability refers to the skill set and expertise of the trustee in some specific field. Benevolence refers to the extent to which the trustee will satisfy the behavioural expectations and avoid opportunistic behaviour. Finally, integrity pertains to the tendency of adhering to the agreed terms and conditions.

Reputation information can be collected from two different sources: direct interactions, and through third-party reputation networks. In the case of previously unknown entities, the third-party reputation networks are the primary source of reputation information. First-hand experiences, as they become available through direct interactions, are valuable as they can be relied on to be both truthful and apply well to the specific trustor's context. A combination of both sources can be used to reason about little-known entities. The assumption of behavioural consistency that any reputation-based predictions rely on can be broken by the trustee at any time; a key measure of the quality of reputation information is whether the actor is also known to behave consistently [4], while the reputation tracking itself also discourages misbehaviour [13].

Contextual Criteria: Contextual criteria comprise factors that vary from situation to situation. They might change even in the presence of the same trustor and trustee. We categorize the key contextual criteria as pertaining to system trust, user interface aiding decision making, and external environmental conditions.

The concept of system trust was introduced by McKnight et al. [39], [42], who proposed it to consist of two components: structural assurances and situational normality. Another component of facilitating factors was later added by Pavlou [43]. Structural assurances consist of legal and governmental impersonal structures, such as legal contracts, safety nets, legal regulations, guarantees and insurances [39], [30]. Structural assurances prove to be a reliable aid for the establishment of trustee trustworthiness during the trust decision making. They are especially useful for trust establishment in the initial stage of trust development, when collaborating with previously unknown or little known entities [30]. Situational normality reflects the trustor's belief that the situation requiring trust decision making is safe and positive for attaining the desired gains [39], [30]. Structural assurances and trustor criteria contribute to the formulation of the trustor's belief. Facilitating factors refer to non-governmental structures, such as shared standards, protocols, relationships, goals or beliefs, which lead to the formulation of a positive perception about the integrity and adherence of the trustee [43].

The user interface acts as an important tool for enabling human trust decision making, as it is responsible for presenting the required information to the users. Our research aims at providing user interfaces to support a range of trust-decision-related tasks, and towards this goal we have built an initial version of the trust decision expert tool that will be providing a user interface for semi-automated trust decisions in Pilarcos (see Figure 1). A literature review reveals the existence of interfaces for making human queries for the existing trust management systems of ECOLEAD [27] and TrustCoM [26] as well. During the initial stage, navigational ease, user-friendly interface, clarity, accuracy and reduced error rate have a positive impact on trust decision making [44], [45]. On the other hand, interactivity, usefulness, accurate transactions together with zero error rates are dominant during the committed stage [44], [46].

External environmental factors refer to the set of social, economic and technological aspects influencing trust decision making. For example, recession could be a major environmental factor affecting trust decisions. The external environmental factors are independent of the trustor, trustee or the contextual trust decision making criteria.

Collaboration-specific criteria: Collaboration-specific criteria refer to the individual objectives and perspectives of the collaborating enterprises that affect trust decision making [6]. The objectives of the enterprises refer to their pre-established intentions about what they wish to gain from the collaboration. These objectives target the perspective the adopts on the trust development. For example, if an enterprise aims at earning money from the collaboration, then it gives an economic perspective to trust development. The perspective reflects the trustor's viewpoint towards trust establishment [6].

We have identified seven different perspectives on trust establishment and decision making: service [47], [6], [2], [25] organizational [6], [2], [25], social [6], [2], economic [6], [2], psychological [24], [25], behavioural [6], [2], [24], [25] and technological [6], [2].

The service perspective refers to taking into consideration the details of service offers. Service offers are made by enterprises who are willing to collaborate [2], [24]. The organizational perspective is made out of enterprise characteristics such as its size and setup [6], [2]. The social perspective for trust decision making targets the outcomes of the interaction of the enterprise with its external environment. One example could be the activities and offerings made by the enterprise towards the surrounding society, through the proper consideration of mutually established contracts, security mechanisms and monitoring [24], [6]. The economic perspective to trust decision making comprises of the involvement of monetary risks or possibilities of making monetary incentives in addition to the financial situation of the enterprise [2], [6]. The psychological perspective refers to the intentions of the enterprise [24]. The behavioural perspective concerns the past and present behaviour of the enterprise [2], [24], [6]. Finally, the technological perspective to trust decision making covers the abilities and competencies of the enterprise [2], [6].

B. Comparison of Trust Management Systems

We will now compare the perspectives on human trust decision making discovered in the literature to the three inter-enterprise collaboration management systems presented in Section II-B. The comparison follows the structure of Section III-A. The key points of the comparison are summarized in Table II.

1) Approaches to Trust Development: TrutCoM follows a cyclic approach to trust establishment [24], [25]. The monitoring of the collaboration is performed during the operational phase. Any kind of misbehaviour is dealt with by activating the appropriate clauses in General Virtual Organisation Agreement (GVOA) and Service Level Agreement (SLA). For example, monetary compensation may be required due to service downtime. Furthermore, all the collaborating enterprises are informed about the observed bad behaviour. This immediate incorporation of any detected breach by taking required actions during the operational phase itself makes their approach to trust decision making cyclic in nature.

In ECOLEAD, trust establishment is carried out in a hierarchical manner [6], [28]. The establishment of specific

Table II SUMMARY OF THE COMPARISON BETWEEN TRUSTCOM, ECOLEAD AND PILARCOS [48].

Comparison Criteria			TrustCoM	ECOLEAD	Pilarcos
Trust Criteria	Trustor Criteria	Propensity to Trust	General Virtual Organization Agreements (GVOA) & Service Level Agreement (SLA)	Criterion governing specific trust	Policies and contracts
		Emotions	Not considered in automated functioning but present in user decision making process	Not considered in automated functioning but present in user decision making process	Trust decision expert tool
		Culture	Business Process Model (BPM) defining different roles & interactions between them	Depends in the administrator	Business Network Model (BNM) defining different roles & interactions between them
	Trustee Criteria	Reputation	• Through characteristics of the service mentioned in published service offer • Information about past behavior of the enterprises stored with the Enterprise Network (EN)	• Base trust established when enterprises enter the Virtual Breeding Environment (VBE) through a questionnaire • VBE storing past information about past behavior of enterprises in earlier collaboratios	Experiences gained from earlier collaborations, first hand information from monitors and shared through reputation networks. The reputation information is represented in the form of assets: monetary, reputation, satisfaction & control
	Contextual Criteria	System Trust	GVOA, SLA, communication standards, monitoring, contract negotiation possibilities & avoidance of information transmission	Hierarchical approach to specific trust, EN consists of enterprises certifying base trust & monitoring	Contract, possibility of contract negotiation, existence of interoperability measures & monitoring
		User Interface	Portlets for interaction with user	Portlets for interaction with user	User interface of trust decision expert tool
		Environmental Factors	Social context	Social perspectives	Contextual repository
	Collaboration specific criteria	Perspectives	Psychological, Social & Behavioral	Organizational, Economical, Social, Technological & Behavioral	Services, Economical, Technological, Behavioral & Risk analysis of threatened assets
		Objectives	Shared & Enterprise specific objectives	Shared & Enterprise specific objectives	Shared & Enterprise specific objectives

trust begins with specifying the underlying objectives which are further categorised into a number of perspectives. Each perspective consists of a set of requirements which are, in turn, divided into value measurement scales and constraints. The measurable elements act as the basis for monitoring. Result updating takes place during the termination phase, so that they can be used during the establishment of future inter-enterprise collaborations. The hierarchal nature of trust establishment and the use of monitoring results only in future collaborations indicate the ECOLEAD approach is staged.

The Pilarcos approach to trust establishment, management and decision making is hybrid, i.e., both cyclic and staged by nature [2], [29]. The Pilarcos middleware performs monitoring during the operational phase, using monitors local to each enterprise. As in the case of TrustCoM, the detection of any significant misbehaviour leads to compensation processes, possible partner reorganization and information dissemination among collaboration enterprises being done during the operational phase of the collaboration. In other words, the information from the current collaboration is being applied both within it and for future collaborations. The results of the monitoring constitute local reputation information, which is fed into third-party reputation systems during the collaboration termination phase [13]. The two kinds of trust decision points makes the Pilarcos approach staged as well. On one hand, trust decisions are made on entering into collaborations with enterprises that carry different levels of familiarity from before. Therefore, especially in the case of previously unknown or little known enterprises, this point is characterized by either no or very low trust, which is equivalent to the situation during the initial stage of trust development. On the other hand, decisions are also made when more resources need to be committed into the collaboration; here the decision can be made based on the experiences gained through the direct interaction with the collaborating enterprises.

2) Criteria for Trust Decision Making: As mentioned before, there is no direct involvement of *trustor criteria* in the automated trust decision making process of all three trust management systems. However, they are reflected in the involvement of human-configured machine agents and humans themselves during policy establishment and contract negotiations. The risk attitudes of the trustor have an effect, for example, as policies are configured for decision-making, and contract templates established for automatically negotiated contracts.

Propensity to trust: The mutually agreed contracts and policies among the collaborating enterprises during the negotiation phase reflect the propensity to trust in Pilarcos middleware [49], [50], [3]. The negotiated contract can be understood as an active and distributed agent containing all the meta-information that governs the working of the inter-enterprise collaboration [47]. The General Virtual Organization Agreement (GVOA) and Service Level Agreement (SLA) reflect the existence of propensity to trust in the case of TrustCoM [51], [24], [25]. Similarly, both GVOA and SLA comprise a set of policies that define the legal framework for the operation of the inter-enterprise collaboration as a whole, and specific to all the services provided by collaborating enterprises. For ECOLEAD, mutually negotiated contracts and criteria directing the trust establishment reflect the propensity to trust in the system [52]. In general, the trustor's propensity to trust is reflected in contracts particularly through any pre-established contract templates and negotiation policies. The specific clauses selected for a particular collaboration can be influenced by these criteria as well, for example through requiring an additional trusted third party to mediate transactions.

Emotions: The user interface of the trust decision expert tool that has been designed and implemented as a result of this research brings the direct involvement of emotions into the workings of the Pilarcos middleware. TrustCoM includes an eLearning portal demo, which has a user interface used for scenario demonstration to the users for finding suitable collaboration partners [26]. Similarly, ECOLEAD also has web and mobile portlets for handling human queries related to the search for potential partners [27]. The user interface is the main channel through which the user's emotional state can strongly affect a trust decision.

Culture: In the case of Pilarcos, the Business Network Model (BNM) constitutes the culture affecting the trust decision making process [49]. The BNM contains information regarding the organisation of the inter-enterprise collaboration. It specifies the policies based on the legal and regulatory systems of the strategic business domain under consideration, roles to be played by the different collaborating enterprises and interaction among them. Similarly, the Business Process Model (BPM) represents culture in the TrustCoM framework [24], [25]. The initiator enterprise willing to establish a collaboration defines the BPM, which contains information about business processes, roles to be fulfilled by the collaborating enterprises, and functional requirements of the collaboration. In the case of ECOLEAD's ICT infrastructure, culture is defined by the administrator through a process referred to as business opportunity characterisation [53]. As in the case of other trust management systems, the business opportunity characterisation process outlines the roles which will be played by the different collaborating enterprises.

The *trustee criterion* of reputation is to a degree supported in all three systems. The TrustCoM framework makes use of reputation information for trust establishment during the collaboration establishment and SLA generation in the negotiation phase. The reputation information exists as service quality, time, cost, integrity, consistency, capabilities, benevolence and past experiences. It is obtained from two different sources: service offers made by the enterprises willing to

collaborate, and the Enterprise Network (EN) gathering it through monitoring the collaboration during the operational phase of the collaboration. The ECOLEAD infrastructure also makes use of reputation information taken from two sources: a baseline trust questionnaire, and the Virtual Breeding Environment (VBE) [6]. The trust questionnaire is filled by all the enterprises willing to collaborate through the VBE, and it gathers information such as organisational size, setup, financial stability, reliability and experience of similar previous collaborations. The VBE contains reputation information gathered by monitoring the functioning of the collaboration. In Pilarcos, reputation information is gathered locally by monitors at each service provider during the operational phase, and from external reputation networks. The reputation information is used for making risk calculations regarding the inter-enterprise collaboration under establishment or in operation. Both local and external reputation information is transformed into a uniform format of number of experiences. Each experience is represented in terms of four assets, reflecting what kind of impact, positive or negative, major or minor, the collaboration in question has had on the assets. The asset types considered are monetary, the service's own reputation, its autonomy and security, and satisfaction of the contract and goals of the collaboration. We will return to the Pilarcos information model in the next section.

Contextual criteria are considered in various ways by the three reviewed systems. For system trust, the existence of monitoring, legally binding contracts and their negotiation possibilities act as its most notable forms. However, in addition to these elements all three trust management systems have certain individual factors contributing to this. In the case of Pilarcos, automated interoperability checking is an additional factor constituting system trust. The presence of communication standards and avoidance of information transformation are extra elements reflecting system trust in TrustCoM. Similarly, the systematic and hierarchical approach to trust establishment support system trust in ECOLEAD.

User Interface: The user interface and the information it provides influences the user's trust in the system itself, in addition to facilitating trust establishment and evaluation towards the trustee in a given collaboration situation. All three trust management systems provide user interfaces for trust management, in varied forms. For Pilarcos, we propose a trust decision expert tool whose user interface presents the users with information on risk, reputation, collaboration context and collaboration progress. The proposed trust decision expert tool targets non-routine cases that require human intervention for trust decision making. TrustCoM also supports information-providing user interfaces and realizes their importance [24], [25]. However, their design falls outside the scope of the TrustCoM project. ECOLEAD facilitates possibilities of user interaction via computer-supported collaborative tools such as mails, chat, calendar and notifications, in addition to the simulated partner search service provided by its test prototype [27].

The exact form of influence the interface has on the user is an interesting research problem in itself. Karvonen et al. have found cultural differences in how users experience the trustworthiness of an online service [54]. An earlier research collaboration in the Trustworthy Widget Sharing (WiSh) project studied how private people interact with the user interface when making decisions to download software for mobile phones provided by other users [55], [23], but more research is needed on the more complex user interfaces for trust management in inter-enterprise collaborations.

External Environmental Factors: We find that all three trust management systems indicate implicit or explicit support for external environmental factors. Pilarcos explicitly supports them through using context as one of the parameters for automated trust decision making. The sources providing the information for the context parameter reflect both internal and external environmental factors, encompassing the internal state of the service, the business state of the service provider enterprise, and the state of the collaboration the enterprise is involved in [29], [2]. On the other hand, TrustCoM and ECOLEAD implicitly support external environmental factors. TrustCoM has adopted the notion that "trust is always set within a social context" [24]. We believe that the social context is affected by different external social, technological and economic factors. Similarly, ECOLEAD also considers social factors as one of the perspectives for assessing trustworthiness of the target entity [6].

Collaboration-specific criteria are similar in terms of objectives, while varying in the selection of perspectives among the three systems.

Perspectives: All three trust management systems have their independent and overlapping perspectives to trust decision making. Pilarcos considers service, economical, technological and behavioural perspectives for trust establishment [49], [50], [29], [2]. Similarly, TrustCoM also considers trust decision making from four different perspectives: service, psychological, behavioural and social [24], [25]. ECOLEAD relies on five different perspectives: organisational, social, economical, technological and behavioural [6].

Objectives: As mentioned previously, there are either shared or individual objectives involved in any inter-enterprise collaboration. Therefore, the dimension of objectives applies equally to all three systems. In each of them, the inter-enterprise collaboration have some shared objectives, in addition to which collaborating enterprises have their own individual objectives behind their participation.

IV. Pilarcos Trust Decision Support

In this section, we present the design of the user interface for supporting human interventions on trust decisions. The

presented user interface extends the Pilarcos trust management system.

A. Pilarcos Trust Management System

In Pilarcos, local trust decisions are made by agents within the enterprise, on behalf of a single service provided by it. The decisions are based on a combination of local, private and shared information, and both the decision policies and policies for collecting and processing the relevant information are specific to a service. Decisions are triggered when a new collaboration is joined, or at points where significant new resources are committed in an ongoing collaboration. The goal of the decisions is to protect the enterprise's assets through evaluating whether the risks and benefits of the given collaboration are in balance.

The Pilarcos trust management system makes automated trust decisions based on seven different parameters: trustor, trustee, action, risk, reputation, importance and context [2]. As discussed previously, the trustor and trustee are business services operating within their respective enterprises. The action represents a collaboration task that needs to be performed, involving a commitment of the trustor's resources. The risk and reputation factors are closely connected: risk estimates present probabilities of different outcomes calculated based on reputation information, which in turn is stored as experience counters of different observed outcomes so far. These experiences are gathered both directly through local monitoring, and as shared information through reputation networks [2]. While shared reputation information may be erroneous and is therefore locally evaluated for credibility, it provides a valuable extension to the first-hand information particularly in the case of actors that are not previously known, and actors who have recently changed their behaviour.

The representation of the experience information determines how it can be used. The classical value imbalance problem [8], for example, stems from experiences only measuring the positive or negative outcome of an action: four small-value transactions weigh four times more than a single large-value transaction. In some cases the costs and benefits of a collaboration cannot easily be measured in monetary terms: a collaborator may pay an agreed-upon service fee, but then violate the contract terms or overload the service to make it fail for other users, causing lost business elsewhere.

In order to balance between better capturing complex outcomes and supporting automated real-time processing of the experience information, Pilarcos represents the effects of the collaboration outcomes on the enterprise, or the business service more specifically. We have applied the idea of asset-based risk analysis to form our model, and introduce four high-level asset classes: monetary, reputation, control and satisfaction [2]. The outcome effects are then represented on the scale of large negative effect, slight negative effect, no effect, slight positive effect and large positive effect that the action or collaboration task has been observed to have on that asset. This condensation of possible outcomes to a set of categories for affected assets improves the reusability and interoperability of reputation information across different enterprises; for example the monetary value of reputation gains is not universal, but highly subjective and even time-dependent.

The monetary asset denotes any resources that can be represented in monetary terms. The reputation asset reflects the trustor's own public relations, appearance in the media, and attitudes of their customers and partners towards them [2]; in contrast, the reputation information discussed above concerns the past behaviour of the trustee. The need for security, privacy and other aspects related to autonomy are represented by the control asset. This amalgam category aims to capture general threats that do not directly translate to monetary or reputation terms, such as threats towards the trust management system itself, service availability, or capacity to enter into new beneficial collaborations coming up during the given deal. Lastly, the degree of fulfillment of the trustor's expectations by the trustee is represented by the satisfaction asset: it is used to measure whether the trustee tends to respect its agreements [2]. While contract satisfaction by itself is not generally considered an asset in the sense of organizational risk analysis, it is such a central part of reputation information that most systems ignore all other dimensions of experience. An organization can decide whether it enters into a given contract with another organization, but it cannot control whether the counterparty follows the contract.

The reputation counters of observed outcomes are converted into a risk estimate by transforming the absolute numbers into ratios: essentially, 5 major positive experiences out of 10 total experiences translates into a probability of 50% of a major positive outcome for that asset. Relevant adjustments are made to accommodate low-stake actions that cannot have a large monetary effect, for example, and credibility-based weighting between local and shared reputation information.

The risk estimate is compared to risk tolerance formulae to determine the outcome of the decision. Risk tolerance checks may be adjusted automatically according to the strategic importance specified for the action; this essentially represents the known benefits of a positive decision, such as not having to compensate other collaborators due to withdrawal from the collaboration during its operation. In the automated trust decision making process, the different factors are also subject to change by the context parameter, which manifests as conditional filters, or modifiers, of the data to allow temporary situational adjustments in the system. One example of a context modifier is insurance, which can apply to all actions in one specific collaboration and either reduce or altogether eliminate the monetary risk involved. The information model and policy options are

discussed in more detail in earlier work [2], [13].

The information and process model for trust management forms the basis for automation as well as the information that can be shown to the user to support a human intervention in a situation where an automated decision cannot be made. Factors to consider in handling human intervention for semi-automated trust decisions include the phenomenon of human trust decision making and information requirements of the human users, which we have discussed in the previous sections. Further into the design phase, additional factors to consider include the appropriate way of presenting the information, and reducing the frequency of calls for human intervention in the future, to ensure that the efficiency of collaboration management is maintained. We discuss these factors in the next section.

B. User Interface

The user interface design of the proposed trust decision expert tool was gained from Nielsen's usability principles for designing user interfaces [44] and different cognitive strategies: Cognitive Fit Theory, Cognitive Load Theory, Unified Theory of Acceptance and Use of Technology and Technology Acceptance Model [56], [57], [58].

In accordance with the Pilarcos trust information model, the user interface presents information about risk, reputation, goals affecting the importance of the action, and context [48]. Within its main information views, it presents further details on the credibility of the information, behavioural changes that can affect the validity of the reputation information, assets endangered according to the risk tolerance comparison, and the progress status of the collaboration when trust decisions need to be made during the operational phase of the collaboration. The information is presented as a combination of textual and graphical formats. Figure 1 shows an example risk view of the user interface; the other major views of reputation, context and progress information are minimized in the screenshot.

On the top, the user interface presents the goals of the inter-enterprise collaboration, such as earning money, gaining experience or building reputation. The importance of the goals that the enterprise has set for the collaboration encourage a positive trust decision. In addition to the goals, the deadline for making the trust decision is prominently shown. Both these information elements and their placement promote transparency.

The risk view presented in the figure shows the produced risk estimate, represented in the form of probabilities of different outcomes. These outcomes for different assets follow the trust information model of Pilarcos, as described in the previous section. The four asset classes correspond to four graphs in the risk information view. The effects on different assets are independent of each other. In the figure, the probability of a slight negative effect on the monetary asset is 0.35, and the probability of a slight reputation gain is 0.4, for example, but it is entirely possible for the collaborator to both cost money and cause negative publicity. When forming a risk estimate, the total probability of 1 is divided between different outcomes towards a specific asset based on the experience information and context filters.

The risk information can be studied through two different views: collaborative and enterprise view. The collaborative view presents collective risk probabilities for the collaboration as a whole; it reflects the fact that even though a trust decision is generally made concerning a one-on-one interaction within a larger collaboration, other participants in the collaboration may have a strong influence on the eventual outcome of the action. A manager may consider placing an order to a generally reliable contractor, for example, yet decide against the plan due to not being able to trust its proposed subcontractors for this collaboration. In contrast, the enterprise view provides information about the risk posed by the single trustee individually. The current version of the trust decision expert tool presents the collaborative and enterprise view in the exact same format, as shown in Figure 1. As an item of future work, we study ways of visualizing an overview of multiple collaborators' reputation and risk information individually to avoid the information loss of automatically merging them into one, such as by weighted averages or selecting the worst case.

The reputation view provides background information to the risk estimate. Reputation information is presented as graphs, as shown in Figure 2. It consists of experiences, which reflect the past and present behaviour of the trustee on the same outcome scale as the risk information. The view also shows the estimated credibility of the shared reputation information, and presents whether the trustee's behaviour has been consistent or not, which may have an impact on the validity of the available reputation information.

Behavioral consistency is expressed through the number of times the system has detected a change in the actor's behaviour [4], and by showing both the overall experiences and the experiences based on the current period of consistent behaviour. Each period of internally consistent behaviour is referred to as a reputation epoch, and the number of epochs reflects the number of times an inconsistency has been observed. In this view, the total number of experiences observed on the trustee is 37, and they are divided into two reputation epochs. In other words, after the first 8 experiences, the trustee's behaviour pattern changed enough that a new reputation epoch was started. The current reputation epoch has lasted for the 29 most recent experiences, 17 (8+9) of which are negative, 7 (4+3) positive and 5 neutral.

The underlying system supports unknown outcomes in experiences as well, due to an implementation convention that each stored experience item contains an outcome value for all four assets. We do not show any unknown outcomes in this version of the user interface, however. Instead, the total number of experiences for, say, the control asset could

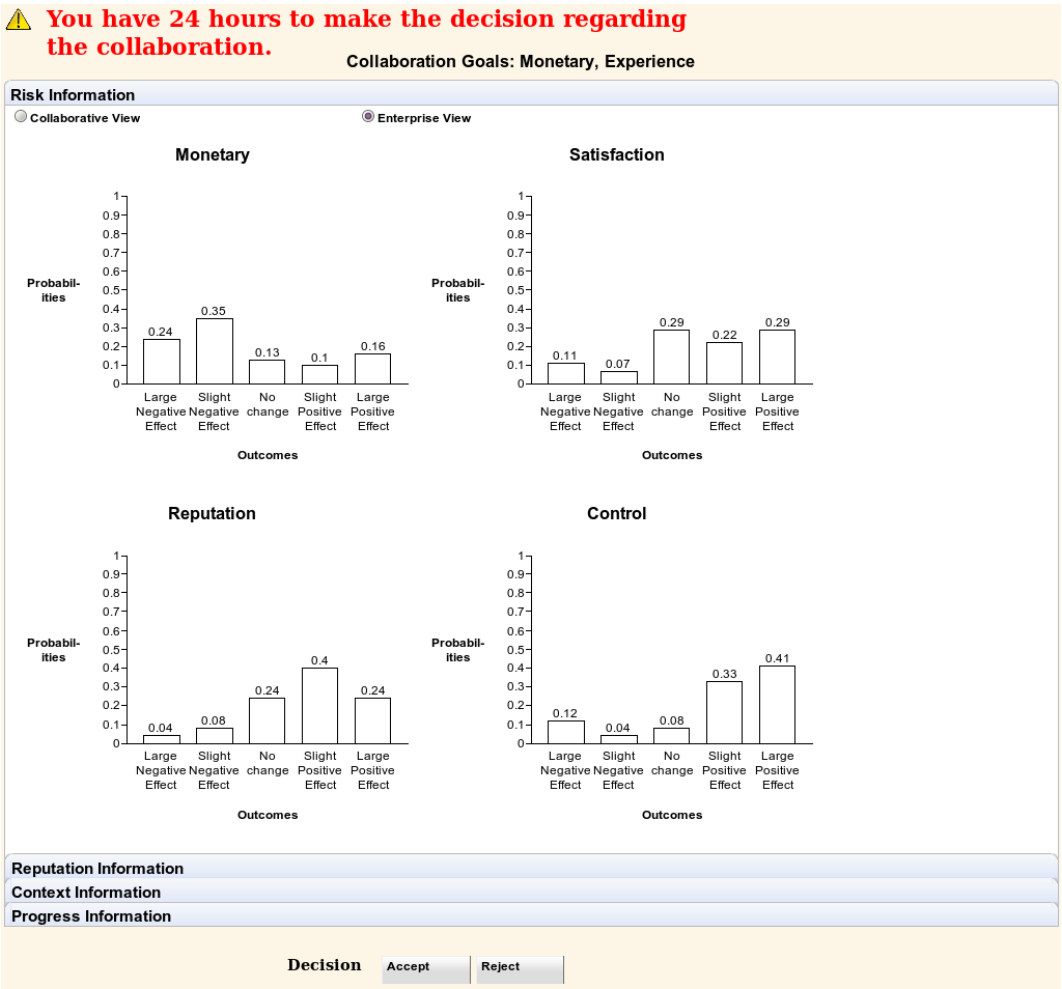


Figure 1. Risk information view of the trust decision expert tool [1].

be lower than the number of experiences for the monetary asset, if the outcome could not be measured in some cases or a reputation source only provided experiences from the monetary and satisfaction points of view, for example.

In Figure 2, the overall view credibility for reputation from the monetary point of view (0.9) is higher than that of the current reputation epoch (0.5), which implies that the recent experiences have come from lower-credibility sources, while the early experiences would be first-hand or from equally trustworthy sources. The exact calculation formula for derived factors such as the view credibility depends on the specific policies in use. To support the repeatability of the user experiments, we have manually configured a fixed output shown to all the users, so they do not directly reflect specific policies in the system.

Finally, the view uses colours to indicate the assets for which the risk estimate is not within the automatically acceptable risk tolerance bounds, as the actor's reputation information for these specific assets may be of particular interest. This information is communicated on the reputation view, where the user interface element doubles as a means to look at the reputation information of a specific asset. The red colour for monetary asset means that the high number of negative experiences gained on the actor, when measuring the monetary effects that collaborating with it has had, translates to a monetary risk estimate that is not considered acceptable for an automated decision. As with risk estimates, the collaborative view of reputation is identical in format to the enterprise view, which means that the reputations of the individual participants must be flattened into a single collection by weighted averages, for example. A way to visualize a collaboration-wide overview that retains relevant information better is an interesting item of future work.

In automated trust decisions in Pilarcos, risk estimates are compared against risk tolerance, which is in turn based on the strategic importance of the action at hand. While the importance is represented through the goals of the collaboration, and tolerance constraints are partially visible

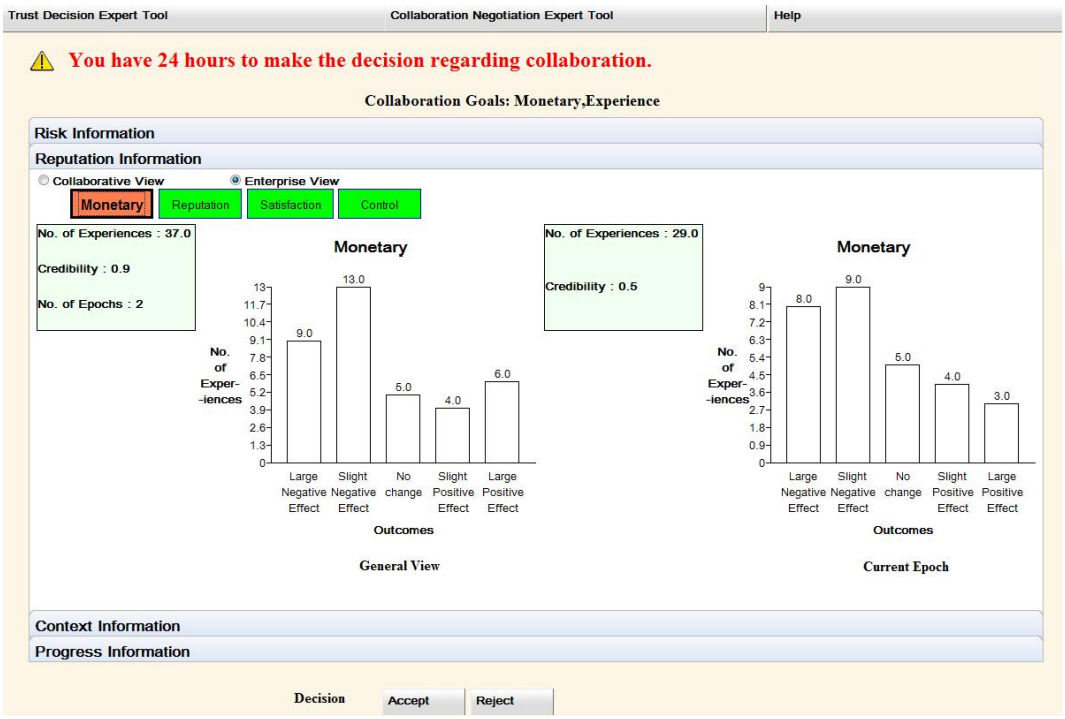


Figure 2. Reputation information view of the trust decision expert tool [48].

through the assets shown not to be within limits, in manual trust decisions the human user is responsible for analyzing and setting the actual risk tolerance limits for the decision.

In the expert tool, the context information view presents the currently active context items to the human users through simple textual phrases, such as “Enterprise A is an important strategic partner” or “The current collaboration is covered by insurance”. This information is collected from the descriptive metadata of any active context filters; we do not consider the formulas of the context filters themselves to be particularly informative to the user.

Finally, the progress information view of the expert tool supports trust decisions on ongoing collaborations. The view presents the progress of the collaboration in graphical format, visualizing the tasks completed by different partners. The views not shown in this paper can be found in the accompanying technical report [48].

The eventual user decision is either to accept and approve the action or reject it, which generally results in a withdrawal from the collaboration. In addition, the tool will ask the user to provide a scope for the trust decision: whether it applies for the remainder of the contract, or for a given time period, or for this specific decision only. This helps reduce the frequency of requests for human intervention, as further decisions needed within the set scope can be automated. The scope information is stored as a context filter, which overrides the risk tolerance formulae appropriately to automatic rejection or acceptance for future decisions within the given scope.

C. User Evaluation

The usability of the trust decision making tools is an important part of enabling user involvement in trust decision making for inter-enterprise collaborations. The tools will not be deployed if are too difficult to use, cannot reflect the user’s information needs, or are conceptually incompatible with human trust decision making processes. We have implemented the trust decision expert tool as a part of enabling user involvement in the Pilarcos collaboration management system, and present here the user study setup and initial results of the user evaluations of the expert tool [48].

We have evaluated the trust decision expert tool interface from four points of view: (i) sufficiency of the presented information, (ii) usability, (iii) user performance and (iv) quality. All five participants in the initial evaluation are researchers in the field of computer science, and one of them was somewhat familiar with the underlying Pilarcos trust management software specifically from before. The main objective behind recruiting technically savvy participants was to gather feedback from users who are representative of the actual target user base of the Pilarcos expert tool. The user study took around one hour per user.

The user studies were conducted in three phases: introduction, solving test tasks and debriefing. During the introduction phase, test participants were first explained the purpose of the study, study setup and details regarding task

performance. Following this, they were informed about the code of ethics and asked to sign the permission form. After the introduction by the moderator, the test participants were presented with the following test scenario:

"You are running an enterprise named 'Quick Service,' which provides online logistic services within Europe. Your enterprise is involved in collaborating online with other enterprises throughout the world. You are using the Pilarcos middleware for managing your online collaborations. Usually, Pilarcos middleware makes automated decisions regarding your enterprise's participation in the online collaborations, but now you have received an email, containing a link, asking you to make a decision regarding your continuation in an ongoing collaboration."

Based on the test scenario, the participants were asked to write down their expectations about information that they would like to have for making trust decisions in such a situation. After the introductory phase, the test participants were allowed to study the user interface to familiarize themselves with it. During the second phase of the user evaluation, the test participants were presented with a set of tasks they are asked to perform using the user interface, one by one. An example of a test task is as follows:

"After reading the email, you already started thinking about the assets that might be endangered by further participating in the collaboration. You figure out money is the most important asset for your enterprise. You decide to find out the risks that the collaboration poses economically to your enterprise."

Each task was further divided into three to four sub-tasks which involved finding required information from the user interface. The sub-tasks, handled one at a time, were provided on paper slips which consisted of a question statement and multiple choices where the participants could mark their answers using a pen or pencil. The decision on presenting the sub-tasks on pieces of paper was made in order to reduce moderator influence. Furthermore, they enabled concluding the test at any time when desired by the test participant, without making them feel uncomfortable for not completing the test.

The test participants were encouraged to think aloud while performing the test tasks. The "think aloud" methodology has been employed to gain insights into the problems and thought process of the participants while they are performing the test tasks [44]. The moderator noted the time taken by test participants to complete each sub-task and comments made while performing it. The completion of each task was followed by a short questionnaire capturing the real-time experience of the participants after each task.

Finally, during the debriefing phase, the test participants were asked to fill in a post-questionnaire aimed at gathering general experience and impressions about the user interface. The post-questionnaire consisted of objective type questions gathering feedback using a five-point Likert scale, ranging from "strongly agree" to "strongly disagree", in addition to open-ended questions. Example claims included: *"The trust decision expert tool is easy to use." "I think the trust decision expert tool presents all the information needed for decision making." "I think the trust decision expert tool presents the information in formats co-related with the task of decision making."* Example open ended questions included: *"In the trust decision expert tool I liked..."* and *"In the trust decision expert tool, I think the following information is missing..."*

The user evaluations have been made with five test participants, which has provided us with useful feedback on further developing the tool. In order to draw any broader conclusions about the usability of the trust decision expert tool, a larger user experiment is necessary. However, as the tool development work is ongoing, the smaller experiment supports particularly the discovery of any glaring issues that should be addressed in the next version of the prototype. We have planned and worked on related decision elements as separate projects that are yet to be bundled together to a coherent set of user interfaces; this bundled whole would then form a natural target for larger-scale testing in future work.

The first point of view evaluated the sufficiency of the information presented to the human users for trust decision making. As previously mentioned, the user interface presents risk, reputation, context, collaboration progress status, goals and credibility information for trust decision making. Table III shows the user rating of the user interface in terms of the sufficiency of the presented information. Based on the analysis of the debriefing phase and participant comments while performing the tasks, we believe the probable reason for disagreement might be the absence of some relevant information, such as the value of the contract in terms of possible monetary profits to the enterprise. Another suggestion for enhancing the available information concerned a more detailed representation of the collaboration progress, relating it to the underlying business process model instead of simple milestones.

The second point of view of usability evaluated how easy the user interface was to use, in terms of ease of finding the presented information, clarity and existence of correlation between the information presentation formats and tasks to be performed. Missing or unclear information were again a likely cause for critique here. For example, the test participants were unclear about the ontological meaning of the assets. Furthermore, some users suggested that a summarized and concise view of the information already presented should be added, including, for example, small textual sentences such as "you have 63% probability of gaining monetary benefits".

The third point of view of the user performance evaluated the user interface in terms of the success rate of task completion, number of errors committed while performing the tasks, and time taken for task performance. The evaluation results

Table III SUMMARY OF THE USERS' OVERALL EVALUATIONS.

Statement	Str. Agr.	Agree	Neutral	Disagree	Str. Dis.
Sufficiency of information	-	3	-	2	-
Ease of finding information	-	3	2	-	-
Clarity fo presentation	-	4	-	1	-
Correlation between information presentation and tasks	2	1	2	-	-
Ease of use	-	3	2	-	-
Confidence of using	-	3	2	-	-
Willingness to use in future	-	4	-	1	-
Feel safe to use	-	-	4	1	-

reveal that the task completion rate is 100% irrespective of accuracy. However, when considering the factor of accuracy, the successful task completion rate is 100% for only two of the participants. It is 93% for two other users, and 78% for one participant. In other words, the error rate is 7%. We suspect the lack of attentive focus while reading the tasks to be the main reason for the existing error rate, because we found the same participants giving correct answers to other similar tasks. Regarding task completion timing, we found that three of the test participants are able to perform 71% of the tasks within seconds, whereas the remaining participants perform respectively 79% and 93% of the tasks in seconds. The times taken here varied from less than 10 seconds seconds to almost a minute depending on the task presented. In contrast, the most laborous of the tasks required up to two and a half minutes (with a mean of 1:28) for the users to find the required information. The time taken is reasonably agreeable considering the novelty of the tool, as none of the participants have ever used any kind of trust decision expert tool before. It does reflect the difficulty of using an expert tool to make manual decisions in complex situations, however: as the available information becomes more intricate, the need to help the user through an intuitive user interface and controlled automation increases.

The fourth point of view of quality aims to evaluate the user satisfaction of using the user interface in terms of ease of use, confidence, willingness to use and perception about security. The evaluations are summarized on the last four lines of Table III. As mentioned previously, insufficient presented information seems to be a likely reason for disagreements or neutral opinions.

In general, we learned that the test participants found the information presentation formats to be easy to read and understand. They also stated they found the user interface to be intuitive, although obviously there is room for improvement.

The user evaluations also resulted in identifying some suggestions in terms of missing information for further improvement of the existing version of the trust decision expert tool. Table IV presents the suggestions or recommendations together with the participant concerns and their priority. The participant concern provides the justification for introducing the proposed change. Priority has been decided based on three factors: effect of the change on the workings of the trust decision expert tool, support provided by the Pilarcos trust management system, and number of participants supporting it. For example, presenting the summarised view of the presented information will affect trust decision making positively and can be supported by Pilarcos. Furthermore, a majority of the test participants expressed their desire of having this kind of information for trust decision making.

V. CONCLUSION AND FUTURE WORK

Inter-enterprise collaborations and social networking have become part of our normal life. To support dynamically evolving open service ecosystems, we need infrastructure to handle interoperability management, collaboration coordination, breach recovery and trust management. Organizational and individual users' decision-making processes and the conceptual basis required by them are not sufficiently supported by automation tools yet. Particularly, the concepts, processes and interfaces for enabling user involvement are missing in many current collaboration management systems. Furthermore, consumers of collaboratively provided services or networked services in general are not sufficiently aware of the trust related threats in this new environment. As a result, they act too trustingly or base their trust decisions on unreliable, or unsuited information. Both in the service provision environments and in the service consumption side it is still early days when it comes to understanding how the complex networks and partners cause threats in trust-requiring activities.

It is essential that research is done on trust management for inter-enterprise collaborations where there is no joint source for information on who to trust. We need to find out how the human or organizational decision-making processes work and what kind of information humans can perceive and judge situations with. We need to find out how trust decisions can be supported with automation, and how that automation or trust decision interfacing can be kept secure. Furthermore, trust decisions often take place side by side with strategical

Table IV SUGGESTIONS, PARTICIPANT CONCERN AND PRIORITY REGARDING RECOMMENDATIONS FOR TRUST DECISION EXPERT TOOL.

Suggestion	Participant Concern	Priority
Summary/Analysis of the presented information	Test participants are concerned about analyzing the presented information. The summary can give them more confidence and enhance clarity.	High
Information about other collaboration alternatives	Test participants are anxious to know other collaboration possibilities. In the case of Pilarcos, this need will be addressed by a collaboration negotiation expert tool in planning.	Low
Previous decision history	Test participants are concerned about the previous decisions made automatically or manually on behalf of the enterprise. This will enhance clarity and confidence in decision making.	Medium
Ontological explanation of presented factors	Test participants are particularly concerned about the exact meaning of the presented information. This will enhance confidence, clarity and perceptions about security.	High
Simulating effects of different policies	Test participants desire to analyze the effects of simulating different policies on presented value. Pilarcos supports only one simulation at the time, however. Availability will enhance confidence, clarity and perception about safety.	Medium
Proper business process representation	Test participants are concerned about current used format for presenting progress information. Good information formats will promote clarity, confidence and safety perception.	High

choices and privacy checks, as in the case of Pilarcos. Thus the larger problem is how to design a human-perceivable process to interact with all these facilities.

In this paper, we have analyzed some of the inter-enterprise collaboration management systems with their trust management approaches. Interfacing with human and organisational users on the collaboration and service aspects is not simply a technical user interface problem, but requires that the systems in their entirety have been built in a way that respects the human and organisational concepts, and the processes of perceiving and making decisions. In contrast to the other approaches, we are studying reputation systems based on contract fulfilment, as subjective experience reports based on pure opinion are problematic from the point of view of reputation system robustness. If there is no way to punish unfair experience reports, the incentives to misuse reputation systems threatens to leave them unusable in practice. Contractually regulated experience reporting improves the quality of the input information to the trust management system.

The contribution of this paper lies in the study and application of the human processes meeting the trust decision support systems. First, the literature study gave us insight on the human perspectives on decision making. Second, we focused on understanding the user interface role and criteria. The systems must approach users as autonomous decision-makers that need both automated support for routine tasks in routine situations, and intervention possibilities in less clear circumstances. Furthermore, an individual may be acting on behalf of an organisation, or at least bound by a hierarchy of regulations in his or her use case context. Third, we applied the gained knowledge to a first-cut user interface design. The small study on the usability of that indicates that our plans on a larger interface that captures decision aspects on privacy, business network model and enterprise strategies as well are justified, and the connectivity of those aspects is indeed expected by the users. The focal points to be presented include i) clear summaries that can be expanded into more detailed information as needed, ii) communication of the ontological meaning of presented decision factors, and iii) access to information about the proposed or existing collaboration as a whole, such as the progress of the collaborative business process. It is also particularly beneficial to represent and simulate the effects of configuring and reconfiguring a collaboration, such as partner changes and different business network model selection.

Our future work include research and development of interfacing processes and user interfaces for collaboration contracts, agents managing collaborations in organisations, and for ecosystems within which the inter-enterprise collaborations are governed. In addition, developing metrics for observing service behaviour would benefit both the reputation-based trust decisions and the re-engineering of collaborations and service portfolios of enterprises.

In larger scale, the future development should include education on trust issues for awareness rising in consumers to require proper management facilities for trust and privacy aspects of the networked services they use in daily life. In addition, standards for distributing trust and reputation related information should be developed with sufficiently wide scope of system models in mind, in order to cover not only client-to-system trust, or client-to-server trust, but true collaborative networks too.

ACKNOWLEDGMENT

We would like to thank our anonymous reviewers for their helpful comments. The research leading to this publication

was conducted as a part of the CINCO group at the University of Helsinki, Department of Computer Science. It has been funded by the Academy of Finland through the Trusted Business Transactions (TBT) project.

REFERENCES

[1] P. Kaur, S. Ruohomaa, and L. Kutvonen, "User interface for trust decision making in inter-enterprise collaborations," in *Proceedings of the Fifth International Conference on Advances in Computer-Human Interactions (ACHI 2012)*. Valencia, Spain: IARIA, Jan. 2012, pp. 122–127, Best paper award. [Online]. Available: http://www.thinkmind.org/download.php?articleid=achi_2012_5_30_20133

[2] S. Ruohomaa and L. Kutvonen, "Trust and distrust in adaptive inter-enterprise collaboration management," *Journal of Theoretical and Applied Electronic Commerce Research, Special Issue on Trust and Trust Management*, vol. 5, no. 2, pp. 118–136, Aug. 2010. [Online]. Available: http://www.jtaer.com/aug2010/ruohomaa_kutvonen_p7.pdf

[3] L. Kutvonen, T. Ruokolainen, S. Ruohomaa, and J. Metso, "Service-oriented middleware for managing inter-enterprise collaborations," in *Global Implications of Modern Enterprise Information Systems: Technologies and Applications*, ser. Advances in Enterprise Information Systems (AEIS). IGI Global, Dec. 2008, pp. 209–241. [Online]. Available: http://www.igi-global.com/reference/details.asp?id=9648

[4] S. Ruohomaa, A. Hankalahti, and L. Kutvonen, "Detecting and reacting to changes in reputation flows," in *Trust Management V*, ser. IFIP Advances in Information and Communication Technology, vol. 358, Copenhagen, Denmark, Jun. 2011, pp. 19–34. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-22200-9_5

[5] P. Kaur and S. Ruohomaa, "Human intervention on trust decisions for inter-enterprise collaboration," in *Post-Proceedings of the EDOC2011 PhD symposium*, ser. Department of Computer Science Series of Publications B, vol. B-2011-1. University of Helsinki, Department of Computer Science, Aug. 2011. [Online]. Available: http://www.cs.helsinki.fi/group/cinco/publications/public_pdfs/kaur11human.pdf

[6] S. S. Msanijla, H. Afsarmanesh, J. Hodik, M. Rehàk, and L. M. Camarinha-Matos, "ECOLEAD deliverable D21.4b: Creating and supporting trust culture in VBEs," EC Information Society, Tech. Rep., Mar. 2006. [Online]. Available: http://www.ve-forum.org/projects/284/Deliverables/D21.4b_Final.pdf

[7] H. van der Heijden, T. Verhagen, and M. Creemers, "Understanding online purchase intentions: contributions from technology and trust perspectives," *Eur. J. Inf. Syst.*, vol. 12, no. 1, pp. 41–48, March 2003. [Online]. Available: http://portal.acm.org/citation.cfm?id=965200.965205

[8] Y. Yao, S. Ruohomaa, and F. Xu, "Addressing common vulnerabilities of reputation systems for electronic commerce," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 7, no. 1, pp. 1–15, Apr. 2012. [Online]. Available: http://www.jtaer.com/statistics/download/download.php?co_id=JTA20120101

[9] A. Jøsang and R. Ismail, "The Beta reputation system," in *Proceedings of the 15h Bled Electronic Commerce Conference*, Bled, Slovenia, Jun. 2002, pp. 324–337. [Online]. Available: http://ecom.fov.uni-mb.si/proceedings.nsf/Proceedings/D9E48B66F32A7DFFC1256E9F00355B37/$File/josang.pdf

[10] A. Jøsang and J. Haller, "Dirichlet reputation systems," in *Proceedings of the Second International Conference on Availability, Reliability and Security (ARES 2007)*. Vienna, Austria: IEEE Computer Society, Apr. 2007, pp. 112–119. [Online]. Available: http://dx.doi.org/10.1109/ARES.2007.71

[11] Y. Wang, C.-W. Hang, and M. P. Singh, "A probabilistic approach for maintaining trust based on evidence," *Journal of Artificial Intelligence Research*, vol. 40, no. 1, pp. 221–267, 2011. [Online]. Available: http://www.jair.org/media/3108/live-3108-5411-jair.pdf

[12] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt, "An integrated trust and reputation model for open multi-agent systems," *Journal of Autonomous Agents and Multi-Agent Systems*, vol. 13, no. 2, pp. 119–154, 2006. [Online]. Available: http://eprints.ecs.soton.ac.uk/12593/

[13] S. Ruohomaa, "The effect of reputation on trust decisions in inter-enterprise collaborations," Ph.D. dissertation, University of Helsinki, Department of Computer Science, May 2012. [Online]. Available: http://urn.fi/URN:ISBN:978-952-10-7912-2

[14] S. Ruohomaa, L. Kutvonen, and E. Koutrouli, "Reputation management survey," in *Proceedings of the 2nd International Conference on Availability, Reliability and Security (ARES 2007)*. Vienna, Austria: IEEE Computer Society, Apr. 2007, pp. 103–111. [Online]. Available: http://dx.doi.org/10.1109/ARES.2007.123

[15] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems: Emerging Issues in Collaborative Commerce*, vol. 43, no. 2, pp. 618–644, Mar. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.dss.2005.05.019

[16] R. Kerr and R. Cohen, "Smart cheaters do prosper: Defeating trust and reputation systems," in *Proceedings of the 8th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009)*, vol. 2. Budapest, Hungary: ACM, May 2009, pp. 993–1000. [Online]. Available: http://dl.acm.org/citation.cfm?id=1558151

[17] V. Cahill *et al.*, "Using trust for secure collaboration in uncertain environments," *Pervasive Computing*, vol. 2, no. 3, pp. 52–61, Aug. 2003. [Online]. Available: http://ieeexplore.ieee.org/iel5/7756/27556/01228527.pdf

[18] "The eBay online marketplace website," 2012, (Accessed 17.12.2012.). [Online]. Available: http://www.ebay.com/

[19] P. Resnick and R. Zeckhauser, "Trust among strangers in internet transactions: Empirical analysis of eBay's reputation system," in *The Economics of the Internet and E-Commerce*, ser. Advances in Applied Microeconomics, vol. 11. Elsevier Science, Amsterdam, 2002, pp. 127–157. [Online]. Available: http://dx.doi.org/10.1016/S0278-0984(02)11030-3

[20] S. Marsh, A. Basu, and N. Dwyer, "Rendering unto Caesar the things that are Caesar's: Complex trust models and human understanding," in *IFIPTM 2012*, ser. IFIP AICT, no. 374, NIT Surat, India, May 2012, pp. 191–200. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-29852-3_13

[21] S. Ries, "Trust in ubiquitous computing," Ph.D. dissertation, TU Darmstadt, Oct. 2009. [Online]. Available: http://tuprints.ulb.tu-darmstadt.de/1948/

[22] B. Fogg, C. Soohoo, D. Danielson, L. Marable, J. Stanford, and E. R. Tauber, "How do people evaluate a web site's credibility?" Stanford Persuasive Technology Lab, Tech. Rep., Oct. 2002. [Online]. Available: http://www.consumerwebwatch.org/dynamic/web-credibility-reports-evaluate-abstract.cfm

[23] K. Karvonen, T. Kilinkaridis, and O. Immonen, "Widsets: A usability study of widget sharing," in *Human-Computer Interaction - INTERACT 2009*, ser. LNCS, vol. 5727, no. II, Uppsala, Sweden, Aug. 2009, pp. 461–464. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-03658-3_50

[24] M. Wilson, D. Chadwick, T. Dimitrakos, J. Doser, A. Arenas, P. Giambiagi *et al.*, "The TrustCoM Framework V0.5," in *6th IFIP Working Conference on Virtual Enterprises (PRO-VE 2005)*, Valencia, Spain, Sep. 2005.

[25] M. Wilson, A. Arenas, D. Chadwick, T. Dimitrakos, J. Doser, P. Giambiagi, D. Golby, C. Geuer-Pollman, J. Haller, K. Stølen *et al.*, "The TrustCoM approach to enforcing agreements between interoperating enterprises," in *Interoperability for Enterprise Software and Applications Conference (I-ESA'06)*, Bordeaux, France, Mar. 2006. [Online]. Available: http://epubs.cclrc.ac.uk/bitstream/898/Trustcom_Interoperability_France.pdf

[26] R. Ratti, M. del Mar Rodrigo Castro, C. A. Ferrandiz, S. Mores, R. Rabelo, R. J. T. Junior, and P. Gibert, "TrustCoM project final report," European Commission, Tech. Rep., Apr. 2007.

[27] R. J. Rabelo, S. Gusmeroli, C. Arana, and T. Nagellen, "The ECOLEAD ICT infrastructure for collaborative networked organizations," in *Network-Centric Collaboration and Supporting Frameworks. IFIP TC 5 WG 5.5, Seventh IFIP Working Conference on Virtual Enterprises*, vol. 224. Helsinki, Finland: Springer, Sep. 2006, pp. 451–460. [Online]. Available: http://dx.doi.org/10.1007/978-0-387-38269-2_47

[28] S. S. Msanjila and H. Afsarmanesh, "HICI: An approach for identifying trust elements — the case of technological trust perspective in VBEs," in *Proceedings of the Second International Conference on Availability, Reliability and Security (ARES 2007)*. Vienna, Austria: IEEE Computer Society, Apr. 2007, pp. 757–764. [Online]. Available: http://dx.doi.org/10.1109/ARES.2007.94

[29] S. Ruohomaa and L. Kutvonen, "Making multi-dimensional trust decisions on inter-enterprise collaborations," in *Proceedings of the Third International Conference on Availability, Security and Reliability (ARES 2008)*. Barcelona, Spain: IEEE Computer Society, Mar. 2008, pp. 873–880. [Online]. Available: http://dx.doi.org/10.1109/ARES.2007.123

[30] X. Zhang and Q. Zhang, "Online trust forming mechanism: approaches and an integrated model," in *Proceedings of the 7th international conference on Electronic commerce*, ser. ICEC '05. Xi'an, China: ACM, Aug. 2005, pp. 201–209. [Online]. Available: http://doi.acm.org/10.1145/1089551.1089591

[31] R. Fung and M. Lee, "EC-Trust (Trust in electronic commerce): Exploring the antecedent factors," in *Americas Conference on Information Systems (AMCIS) Proceedings*, Milwaukee, Wisconsin, USA, Aug. 1999, paper 179. [Online]. Available: http://aisel.aisnet.org/amcis1999/179

[32] T. Deelmann and P. Loos, "Trust economy: Aspects of reputation and trust building for SMEs in e-business," in *Americas Conference on Information Systems (AMCIS) Proceedings*, Dallas, Texas, USA, Aug. 2002, paper 302.

[33] D. L. Shapiro, B. H. Sheppard, and L. Cheraskin, "Business on a handshake," *Negotiation Journal*, vol. 8, no. 4, pp. 365–377, Oct. 1992. [Online]. Available: http://dx.doi.org/10.1111/j.1571-9979.1992.tb00679.x

[34] S. Ba, "Establishing online trust through a community responsibility system," *Decision Support Systems*, vol. 31, no. 3, pp. 323 – 336, 2001. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923600001445

[35] E. Kim and S. Tadisina, "Customers' initial trust in e-businesses: How to measure customers' initial trust," in *Americas Conference on Information Systems (AMCIS) Proceedings*, Tampa, Florida, USA, Aug. 2003, paper 5.

[36] D. H. Mcknight, V. Choudhury, and C. Kacmar, "Trust in e-commerce vendors: a two-stage model," in *International Conference on Information Systems (ICIS'00)*, Brisbane, Queensland, Australia, Dec. 2000, pp. 532–536. [Online]. Available: http://dl.acm.org/citation.cfm?id=359640.359807

[37] J. Salo and H. Karjaluoto, "A conceptual model of trust in the online environment," *Online Information Review*, vol. 31, no. 5, pp. 604–621, 2007.

[38] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," *The Academy of Management Review*, vol. 20, no. 3, pp. 709–734, July 1995.

[39] L. L. Cummings, D. H. McKnight, and N. L. Chervany, "Initial trust formation in new organizational relationships," *The Academy of Management Review*, vol. 23, no. 3, pp. 473–490, July 1998.

[40] J. R. Dunn and M. E. Schweitzer, "Feeling and believing: The influence of emotion on trust," *Journal of Personality and Social Psychology*, vol. 88, no. 5, pp. 736–748, May 2005. [Online]. Available: http://psycnet.apa.org/doi/10.1037/0022-3514.88.5.736

[41] F. D. Schoorman, R. C. Mayer, and J. H. Davis, "An integrative model of organizational trust: Past, present, and future," *Academy of Management Review*, vol. 32, no. 2, pp. 344–354, 2007.

[42] D. H. McKnight and N. L. Chervany, "The meanings of trust," University of Minnesota, MIS Research Center, Tech. Rep., 1996. [Online]. Available: http://misrc.umn.edu/workingpapers/fullPapers/1996/9604_040100.pdf

[43] P. A. Pavlou, "Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model," *International Journal of Electronic Commerce*, vol. 7, no. 3, pp. 101–134, Apr. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=1288216.1288221

[44] J. Nielsen, *Usability Engineering*. Boston: M.A. Academic Press, 1993.

[45] G. Häubl and V. Trifts, "Consumer decision making in online shopping environments: The effects of interactive decision aids," *Marketing Science*, vol. 19, no. 1, pp. 4–21, 2000. [Online]. Available: http://dx.doi.org/10.1287/mksc.19.1.4.15178

[46] J. M. Weber, D. Malhotra, and J. K. Murnighan, "Normal acts of irrational trust: Motivated attributions and the trust development process," *Research in Organizational Behavior*, vol. 26, pp. 75–101, 2004, an Annual Series of Analytical Essays and Critical Reviews. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0191308504260038

[47] L. Kutvonen, T. Ruokolainen, and J. Metso, "Interoperability middleware for federated business services in web-Pilarcos," *International Journal of Enterprise Information Systems, Special issue on Interoperability of Enterprise Systems and Applications*, vol. 3, no. 1, pp. 1–21, Jan. 2007. [Online]. Available: http://www.cs.helsinki.fi/group/cinco/publications/public_pdfs/kutvonen06interop.pdf

[48] P. Kaur, "Users' trust decisions on inter-enterprise collaborations," Master's thesis, Aalto University, Computer Science and Engineering; Tech. Rep, University of Helsinki Department of Computer Science, Sep. 2011. [Online]. Available: http://www.cs.helsinki.fi/group/cinco/publications/public_pdfs/kaur11msthesis.pdf

[49] L. Kutvonen, J. Metso, and S. Ruohomaa, "From trading to eCommunity management: Responding to social and contractual challenges," *Information Systems Frontiers (ISF) - Special Issue on Enterprise Services Computing: Evolution and Challenges*, vol. 9, no. 2–3, pp. 181–194, Jul. 2007. [Online]. Available: http://dx.doi.org/10.1007/s10796-007-9031-x

[50] L. Kutvonen, S. Ruohomaa, and J. Metso, "Automating decisions for inter-enterprise collaboration management," in *Pervasive Collaborative Networks. IFIP TC 5 WG 5.5 Ninth Working Conference on Virtual Enterprises, September 8–10, 2008, Poznan, Poland*, ser. IFIP, no. 283. Poznan, Poland: Springer, Sep. 2008, pp. 127–134. [Online]. Available: http://www.cs.helsinki.fi/group/cinco/publications/public_pdfs/kutvonen08automating.pdf

[51] S. Crompton, M. Wilson, A. Arenas, L. Schubert, D. I. Cojocarasu, J. Hu, and P. Robinson, "The TrustCoM general virtual organization agreement component," in *UK e-Science All Hands Meeting, UK Natl e-Science Centre*, Nottingham, UK, Sep. 2007. [Online]. Available: http://www.allhands.org.uk/2007/proceedings/papers/771.pdf

[52] R. Ratti, M. del Mar Rodrigo Castro, C. A. Ferrandiz, S. Mores, R. Rabelo, R. J. Tramontin Junior, and P. Gibert, "Deliverable D61.1c ICT-I Reference Framework (version 3)," European Commission, Tech. Rep., Apr. 2007.

[53] T. Ruokolainen, S. Ruohomaa, and L. Kutvonen, "Solving service ecosystem governance," in *Proceedings of the 15th IEEE International EDOC Conference Workshops*. Helsinki, Finland: IEEE Computer Society, Aug. 2011, pp. 18–25. [Online]. Available: http://dx.doi.org/10.1109/EDOCW.2011.43

[54] K. Karvonen, L. Cardholm, and S. Karlsson, "Designing trust for a universal audience: A multicultural study on the formation of trust in the Internet in the Nordic countries," in *Proceedings of the First International Conference on Universal Access in HCI (UAHCI'2001)*, New Orleans, LA, USA, Aug. 2001. [Online]. Available: http://www.tml.hut.fi/Research/TeSSA/Papers/Karvonen/designing-trust.pdf

[55] Y. Shen, P. Nurmi, S. Ruohomaa, and M. Lehtimäki, "D6.3.2.8: Understanding widget downloading preferences," TIVIT, ICT SHOK Future Internet Programme, Tech. Rep., Mar. 2011. [Online]. Available: http://www.futureinternet.fi/publications/Deliverable_D6.3.2.8.pdf

[56] I. Vessey and D. Galletta, "Cognitive fit: An empirical study of information acquisition," *Information Systems Research*, vol. 2, no. 1, pp. 63–84, 1991. [Online]. Available: http://isr.journal.informs.org/cgi/content/abstract/2/1/63

[57] L. Oshlyansky, P. Cairns, and H. Thimbleby, "Validating the unified theory of acceptance and use of technology (UTAUT) tool cross-culturally," in *Proceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI...but not as we know it - Volume 2*, ser. BCS-HCI '07. Swinton, UK: British Computer Society, Sep. 2007, pp. 83–86.

[58] J. Sweller, "Cognitive load during problem solving: Effects on learning," *Cognitive Science*, vol. 12, no. 2, pp. 257 – 285, 1988. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0364021388900237

A Generic Approach towards Measuring Level of Autonomicity in Adaptive Systems

Thaddeus Eze, Richard Anthony, Alan Soper, and Chris Walshaw
Autonomic Computing Research Group
School of Computing & Mathematical Sciences (CMS)
University of Greenwich, London, United Kingdom
{T.O.Eze, R.J.Anthony, A.J.Soper and C.Walshaw}@gre.ac.uk

***Abstract*— This paper is concerned with setting the groundwork for the introduction of standards for Autonomic Computing, in terms of technologies and the composition of functionalities as well as validation methodologies. This is in line with addressing the lack of universal standards for autonomic (self-managing) systems and design methods used for them despite the increasingly pervasiveness of the technology. There are also significant limitations to the way in which these systems are assessed and validated, with heavy reliance on traditional design-time techniques, despite the highly dynamic behaviour of these systems in dealing with run-time configuration changes and environmental and context changes. These limitations ultimately undermine the trustability of these systems and are barriers to eventual certification. We propose that the first vital step in this chain is to introduce robust techniques by which the systems can be described in universal language, starting with a description of, and means to measure the extent of autonomicity exhibited by a particular system. Existing techniques have mainly qualitatively classified autonomic systems according to some defined levels with no reference to the building blocks (core functionalities) of the systems. In this paper we present a novel and generic technique for measuring the Level of Autonomicity along several dimensions of autonomic system self-* (e.g., self-configuration, self-healing, self-optimisation and self-protection) functionalities. To demonstrate the feasibility and practicability of our approach, a case example of two different scenarios is examined. One example focuses on a specific case approach for LoA measure within a Dynamic Qualitative Sensor Selection scenario. The second example is a deployment of a generic case approach to an envisioned Autonomic Marketing System that has many dimensions of freedom and which is sensitive to a number of contextual volatility.**

Keywords - autonomicity; level of autonomicity; autonomic system; trustworthiness; metrics; autonomic marketing, sensor selection

I. INTRODUCTION

Autonomic Computing (AC) seeks the development of self-managing (or autonomic) systems to address management complexities of systems. The high rate of advancement of autonomic technology and methodologies has seen these systems increasingly deployed across a broad range of application domains yet without universal standards. Also the widening acceptance of Autonomic Systems (AS) is leading to more trust being placed in them with little or no basis for this trust, especially in the face of significant limitations regarding the way in which these systems are validated. The traditional design-time validation techniques fail to address the run-time requirements of AS' environmental and contextual dynamism. These limitations undermine trustability and ultimately impinge on certification. The more this proliferation goes on without these challenges being addressed, the more difficult it gets to introduce standards and eventually achieve certifiable AS. It has therefore become pertinent and timely to address these issues. A vital first step in this course would be standards for the universal description of these systems and a standard technique for measuring Level of Autonomicity (LoA) achieved by these systems –and we have made progress in this area [1]. Standards for AC would be concerned with technologies, composition of functionalities and validation methodologies. By autonomicity we mean the ability of a system to pursue its goal with minimal external interference in the form of configuration or control. Then, the extent of this interference defines autonomicity levels. Now the questions facing the AC community are, for a given system, "How autonomic should a system be?" and "How autonomic is a system and how is this determined?" The two questions address both pre and post system design phases. The first question is of primary importance to the designers of systems where autonomic specification is a critical part of the whole system requirements definition. A good example would be the spaceflight vehicles addressed in [2], where a *level of autonomy assessment tool* was developed to help determine the level of autonomy required for spaceflight vehicles. The second question is in two parts. On the one hand is the need to define systems according to a measure of autonomicity and another is the method and nature of the measure. Addressing this issue is the main thrust of this paper and here we improve on our initial work [1] in this area. Another significant aspect addressed here is the need for a standard way for assessing, comparing and evaluating different systems (with flexibility across many domains) and also in terms of their individual functionalities. Not only do we measure autonomicity but also look at how systems can be evaluated and compared in terms of their autonomic compositions.

Eze *et al* [3] identified that defining LoA is one of the critical stages along the path towards certifiable AS. Along this path also is the need for an appropriate testing methodology that seeks to validate the AS decision-making process. But to know what testing (validation) is appropriate requires knowledge of the system in terms of its extent of autonomicity. Another issue that underpins the need for measuring LoA is that a means of answering the identified questions is also a solution for assessing AS and facilitates a proper understanding of such systems.

Currently, the vast majority of research effort in this direction has progressed in answering the first question ("How autonomic should a system be?") by providing us with scales that describe and analyse autonomy in systems. These

scales, referenced by many researchers, provide fundamental understanding of system autonomy by categorising autonomy according to level of human-machine involvement in decision-making and execution. A naturally upcoming concern with this approach is that high human involvement does not always necessarily translate to low autonomicity and vice versa. Also, most (if not all) of such approaches do not assess ASs based on demonstrated functionalities but on perceived or observed outcomes (performance). Some key works in this area include [2], [4], and [5]. For us, these scales only characterise autonomy levels qualitatively and offer no generic or robust means of quantitatively measuring extent of autonomicity. We would simply say that they are more sufficient for the purposes of proposing an appropriate level of autonomy during the design of a new system.

ISO/IEC 9126-1 standard [6] decomposes overall software product quality into characteristics, sub characteristics (attributes) and associated measures. Adapting this, we define a framework for measuring LoA along several dimensions of AS self-* functionalities. Systems are well-defined by their set of functional capabilities and a measure of these capabilities will form a better representation of the systems. These functional capabilities may be extended to mean, in other systems, characteristics (or attributes) and sub-characteristics (or sub-attributes). While in our initial work [1] we restrict the functionalities to the core functionalities of ASs, the self-CHOP (self-configuration, self-healing, self-optimisation and self-protection) functionalities, in this paper we extend the reach (scope) to cover all possible essential functionalities and identify specific metrics for each of the functionalities. (This allows the approach to be entirely more generic.) The cumulative measure of these metrics defines a LoA. Our method is based on the establishment of a generic technique that can be applied to any application domain. This work is novel as it offers a quantitative measure of LoA in terms of system's functionalities-based description and can be flexibly applied across different application instances. It also opens a new research focus for autonomicity measuring metrics. We believe this is timely because if not addressed we not only run the risk of classifying systems as trusted without basis but also risk losing track and control of these systems as a result of spiraling complexities in terms of technology and methodologies. [7] also raised the concern that if the proliferation of unmanned systems (and by extension ASs) is not checked by putting appropriate measures (or mechanisms) in place that ensure trustworthiness, the systems may ultimately lose acceptance and popularity.

The remainder of this paper is organised as follows: related work is presented in Section II. In Section III, we introduce metrics for measuring autonomicity. Our proposed LoA measure and two case studies are presented in Sections IV and V respectively. Section VI concludes the work.

II. RELATED WORK

The study of AC is now a decade old. However, its rapid advancement has led to a wide range of views on meaning, architecture, and implementations. The criticality of understanding *extent* of autonomicity in defining AC systems has necessitated the need for evaluating these systems. The majority of research in this area has targeted specific application domains with datacentre applications topping the list [8]. Now, to the extent of our research review [8], there is no known (or published) quantitative approach for assessing autonomic systems. There are nonetheless, efforts towards classifying ASs according to *extent of autonomicity* but these efforts have not successfully met the need for assessing autonomic systems. In this section we review some of the proposed (existing) approaches.

One major proposal for classifying ASs according to *extent of autonomicity* (or measuring LoA) is the *scale-based* approach. This approach, based on level of human-machine involvement in decision-making and execution, uses a scale of (1 – *max*) to define a system's LoA where '1', the lower bound, is the lowest autonomic level usually describing a state of least machine involvement in decision-making and '*max'*, the upper bound, is the highest autonomic level describing a state of least human involvement. Prominent in this category of approach are efforts in [2], [4], [9, 27], and [20]. Clough [4] proposes a scale of (1–10) for determining Unmanned Aerial Vehicles' (UAV's) autonomy. Level 1 '*remotely piloted vehicle'* describes the traditional remotely piloted aircraft, while level 10 '*fully autonomous'* describes the ultimate goal of complete autonomy for UAVs. Clough populates the levels between by defining metrics for UAVs. Sheridan [9] also proposes a 10-level scale of autonomic degrees. Unlike Clough's scale, Sheridan's levels 2-4 centre on who makes the decisions (human or machine), while levels 5-9 centre on how to execute decisions. Ryan *et al* [2], in a study to determine the level of autonomy of a particular AS decision-making function, developed an 8-level autonomy assessment tool. The tool ranks each of the OODA (Observe, Orient, Decide and Act) loop functions across Sheridan's proposed scale of autonomy [9]. OODA is a decision-making loop architecture for ASs. The scale's bounds (1 and 8) correspond to complete human and complete machine responsibilities respectively. They first identified the tasks encompassed by each of the functions and then tailored each level of the scale to fit appropriate tasks. The challenge here is ensuring relative consistency in magnitude of change between levels across the functions. The levels are broken into three sections. Levels 1-2 (human is primary, computer is secondary), levels 3-5 (computer and human have similar levels of responsibility), and levels 6-8 (computer is independent of human). To determine the level of autonomy needed to design into a spaceflight vehicle, Ryan *et al* [2] needed a way to map particular functions onto the scale and determine how *autonomous* each function should be. They designed a questionnaire and sent it to system designers, programmers and operators. The questionnaire considered what they call '*factors for determining level of autonomy*', which include level of autonomy *trust limit* and *cost/benefit ratio limit*. This implies that a particular level of autonomy for a function is favoured when a balance is struck between *trust* and *cost/benefit ratio limits*. Ultimately the pertinent question

is "How autonomous should future spaceflight vehicles be?" This is a brilliant technique towards answering the first identified question ("How autonomic should a system be?") IBM's 5 levels of automation [5] describes the extent of automation of the IT and business processes. We consider these to be too narrowly defined and [10] observes that the differentiation between levels is too vague to describe the diversity of self-management, making it difficult to align ASs with those levels [28]. One major concern with the *scale-based* approach is that a system is not necessarily less autonomic when human interferes with its operations and vice versa. Another is the complexity of applying the approach across different application instances (systems) –this is in terms of populating the levels in-between the scales: the differentiation between levels is complex (and can vary significantly depending on who is using the approach) to determine appropriate magnitude for each level. In general the autonomy scale approach is qualitative and does not discriminate between behaviour types. We posit that a more appropriate approach should comprise both qualitative and quantitative (as a way of assigning magnitude or value to the description and classification of systems) measures. These concerns are considered and addressed in our approach.

Hui-Min *et al.* [20] is a government's front for addressing the challenge of classifying the pervasive unmanned systems (UMS) according to their levels of autonomy. [20] alludes that UMS' autonomy cannot be rightly evaluated quantitatively without thorough technical basis and that the development of autonomy levels for unmanned systems must take into account factors like task complexity, human interaction, and environmental difficulty. The product in [20] is Autonomy Levels for Unmanned Systems (ALFUS) Framework which, more specifically, provides the terminology for prescribing and evaluating the level of autonomy that an unmanned system can achieve. The framework, in which the levels of autonomy can be described, addresses the technical aspects of UMS and includes terms and definitions (set of standard terms and definitions that support the autonomy level metrics), detailed model for autonomy levels, summary model for autonomy levels, and guidelines, processes, and use cases. While we accept that autonomicity cannot be correctly evaluated without thorough technical basis, our approach further takes into account key functionalities of ASs rather than individual breakdown of technical operations and operational conditions –a major difference with our work. The work in [20], which is updated in [21], focuses more on standardised categorisation of UMS.

Barber and Martin [11] supposes that in a multi-agent system environment, agent autonomy is measured in terms of a system-wide goal. It proposes a collaborative decision-making algorithm for multi-agent systems. In the proposed algorithm, a plan for achieving the system's goal is decided by the agents. Every agent suggests a complete plan with justification for how to achieve the entire system's goal. Each agent evaluates each suggested plan and determines the value of its justification. Each plan receives an integer number of votes from the deciding agents. The plan with the highest votes becomes the plan for the entire system. The ratio of an agent's number of votes (received for suggested plan) to the total number of votes cast is a measure of that agent's autonomy and the extent of its capability to influence the system. This method, however, does not offer a measure for LoA but gives a valuable description of agents' individual influence in a multi-agent system environment which is useful to our approach: In further evaluating a system, we adapt this formula to determine the rate of individual functionality contribution in our proposed LoA measure (see Section IV B).

Fernando *et al* [12] proposes measures for evaluating the autonomy of software agents. It believes that a measure of autonomy (or any other agent feature) can be determined as a function of well-defined characteristics. Firstly, it identifies the agent autonomy attributes (as *self-control*, *functional independence*, and *evolution capability*) and then defines a set of measures for each of the identified attributes. The agent's LoA is defined by normalising the results of the defined measures using a set of functions. [12] considers autonomicity measure with reference to system's characteristics and attributes. But in that work *'characteristics'* are a broad range of attributes that describe a system which also include features outside the system's core functionalities. Not going into the argument of right/wrong constitution of system attributes (or functionalities), the important aspect to note is the idea of defining a system with respect to its attributes and characteristics. We have adapted this approach in our proposal for autonomic systems but with reference to [core] autonomic self-* functionalities.

III. AUTONOMICITY MEASURING METRICS

In this section, we introduce example metrics for each of the core four functionalities that define autonomicity of AS. Though metrics are application domain dependent, the metrics presented here are generic and serve as examples only. We understand that autonomic functionalities are emergent and these vary (or are defined) according to application instances. The point is that, for any system (whether or not autonomic), there are required functionalities (determined by designers and/or users) which can be measurable by some identified metrics. We present at least one metric for each of the functionalities (using the self-CHOP for example). This is part of a wider (and separate) research focus. This section only focuses on how autonomic metrics can be generated. We also show how metrics can be normalised (see Section IV). We will start with a definition of each CHOP. (For more on these definitions see [13] and [14]).

***Self-Configuring*:** A system is self-configuring when it is able to automate its own installation and setup according to high-level goals. When a new component is introduced into an AS it registers itself so that other components can easily interact with it. The extent of this interoperability I is a measure of self-configuration, measured as the ratio of actual number of components ($n_{i_{actual}}$) to expected number of components ($n_{i_{expected}}$) successfully interacting with the new component after configuration.

$$I = \sum_{1}^{i} \frac{n_{i_{actual}}}{n_{i_{expected}}} \quad (1)$$

Interoperability ratio I measures to what extent a system is distorted by an upgrade. A system is self-configuring to the extent of its ability to curb this distortion. This example can be related to the problem diagnosis system for AS upgrade discussed in [13]. Here an upgrade introduces 5 software modules. The installation regression testers found faulty output in 3 of the new modules. This implies that only 2 modules out of 5 successfully integrated with the system.

***Self-Optimising*:** A system is self-optimising when it is capable of adapting to meet current requirements and also of taking necessary actions to self-adjust to better its performance. Resource management (e.g., load balancing) is an aspect of self-optimisation. An autonomic system is required to be able to learn how to adapt its state to meet new challenges. Also needed is consistent update of the system's knowledge of how to modify its state. State is defined by a set of variables such as current load distribution, CPU utilization, resource usage, etc. The values of these variables are influenced by certain event occurrences like new requirements (e.g., process fluctuations or disruptions). By changing the values of these variables, the event also changes the state of the system. The status of these variables is then updated by a set of executable statements (policies) to meet any new requirement. A typical example would be an autonomic job scheduling system. At first, the job scheduler could assign equal processing time quanta to all systems requiring processing time. The sizes of the time quantum becomes the current state and as events occur (e.g., fluctuations in processing time requirement, disruptions of any kind, etc.), the scheduler is able to adjust the processing time allocation according to priorities specified as policies. In this way the state of the system is updated. But this may lead to erratic tuning (as a result of over or under compensation) causing instability in the system. We define *Stability* as a measure of self-optimisation. If an event leads to erratic behaviour, incoherent results or system is not able to retrace its working state beyond a certain safe margin (a margin within which instability is tolerated), then the system is not effectively self-optimising.

***Self-Healing*:** A system is self-healing when it is able to detect errors or symptoms of potential errors by monitoring its own performance and automatically initiate remediation [15]. Fault tolerance is one aspect of self-healing. It allows the system to continue its operation possibly at a reduced level instead of stopping completely as a result of a part failure. One critical factor here is latency; the amount of time the system takes to detect a problem and then react to it. We define *reaction time* T as a metric for self-healing capability. This is crucial to the reliability of a system. If a change occurs at time t_a and the system is able to detect and work out a new configuration and ready to adapt at time t_b, then (2) defines the reaction time T. (Average is taken instead where variations of T are possible).

$$T = t_b - t_a \quad (2)$$

A case scenario is a stock trading system where time is of paramount importance. The system needs to track changes (e.g., in trade volumes, price, rates etc.) in real time in order to make profitable trading decisions.

Self-Protecting: A system is self-protecting when it is able to detect and protect itself from attacks by automatically configuring and tuning itself to achieve security. It may also be capable of proactively preventing a security breach through its knowledge based on previous occurrences. While self-healing is reactive, self-protecting is proactive. A proactive system, for example, would maintain a kind of log of trends leading to security problems (threats and breaches) and a list of solutions to resolve them (a list of problems and corresponding solutions only applies to self-healing). One major metric here is the ability of the system to prevent security issues based on its experience of past occurrences. For example let's assume $p \in \{p_{ij}\}$ to be true if i^{th} trend leads to j^{th} problem where p_{ij} is a log of all identified trends and corresponding problems. p is a particular instance of trend-problem combination. A self-protecting manager will avoid a situation of same trend leading to the same problem again by blocking the problem, addressing it or preventatively shutting down part of the system. We define *ability to detect repeat events* E as a self-protecting metric. E is a Boolean value (True indicates the manager is able to stop a repeating problem and False otherwise). If we choose two samples of $\{p_{ij}\}$ at different times (t_1 and t_2) then (3) defines E. (Different trends may lead to the same problem but a repeated trend-problem combination indicates a failure of the system to prevent a reoccurrence).

$$E = True\ \forall_{ij} \text{ if } \{p_{ij}\}t_1 \cap \{p_{ij}\}t_2 = \emptyset \quad (3)$$

One typical implementation of this is an antivirus system. Some antivirus systems learn about trends or patterns (signatures) and are able to make decisions based on these to proactively protect a system from an attack. The antivirus is able to stop repeatable patterns. Detecting problem reoccurrence is an active research focus in Autonomic Computing [16].

IV. PROPOSED LOA MEASURE

An AS is defined based on its achievement of the self-* capabilities [15]. In our approach, we define a level of AS in terms of its extent of achieving the identified functionalities. If a system fails to demonstrate at least a certain level of one of the self-* (required for the system in question), the system is said to be non-autonomic. On the other hand, if the system demonstrates a full level of all identified (or required) capabilities, it is said to have achieved full autonomicity (as defined by our proposed scheme). In this section, we present our updated approach towards measuring autonomic systems LoA. In the most part, a mathematical algorithm is used for the proposed approach.

Each functionality is defined by a set of metrics. Each functionality contributes a level of autonomic value which is spread across the set of metrics for that functionality. It then follows that each metric contributes a certain quota of the autonomic value for that functionality. Metrics and functionalities are weighted according to relevance or importance. The cumulative normalisation of the measure of all metrics (for all functionalities) defines a LoA. The need for normalisation of values enables comparison of systems across different implementations. With an ongoing debate on the composition of AS functionalities and the list substantially growing [17, 18], our approach is generic to accommodate evolving functionalities as may be defined by the user. Figure 1 is a pictorial illustration of our approach.

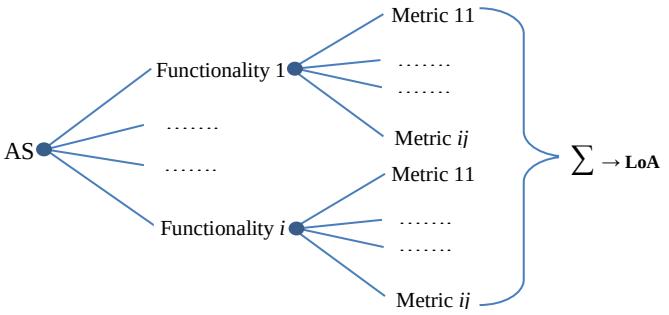


Figure 1: Pictorial illustration of how LoA is achieved by summing the metric autonomic value contributions of all metrics defining all functionalities of a particular AS.

Given that any AS is defined by a number of self-* autonomic functionalities, say n, the mathematical *Combination* expression (4) is the representation of the possible combinations of the functionalities:

$$\sum_{r=1}^{n} {}^{n}C_{r} \rightarrow \#\text{ of possible combinations} \qquad (4)$$

The number of possible combinations indicates the possible functionality compositions of a system where n is the number of functionalities (the self-*) and r is an enumerator of the possible implementation combinations (see the rightmost enumerated values in Figure 2). The functionalities may not be of equal importance to an application domain so combinations indicate what functionality is important to an application domain. And depending on choice of usage, this may be defined as *required* functionalities (in which case r may be equal to n) or demonstrated functionalities (in which case $r \leq n$).

Autonomic functionalities may overlap i.e., are not necessarily orthogonal. For example, a function that primarily achieves self-healing may change internal configuration and thus may also be described as self-configuring. To represent this, we allocate weights to indicate the extent to which a particular algorithm achieves the different functionalities. Further, self-managing actions are not necessarily linear in their operation; i.e., the relationship between a self-tuning parameter change internally and the externally seen effect of the change may be non-linear. In addition, for a given system, one autonomic behaviour e.g., self-healing may have a much more significant effect on system behavior than perhaps self-optimisation which may be more subtle. Such non-linearity in the contribution to LoA is catered for by a combination of weighting and normalisation (see Section IV part C). Weights are applied to reflect the extent of impact one of a particular functionality. Our current technique caters for orthogonality and non-linearity although to some extent these are open challenges that need further addressing.

Table I is a description of notation keys used. To measure the LoA of a system, we require the following:

- Number of functionalities: this is a value indicating the number of functionalities present or required in a particular system – a specific implementation combination of the functionalities.
- Number of metrics: this is the number of identified metrics for the respective functionalities.
- Weighting: weights are assigned to functionalities and metrics according to priority or importance.

TABLE I: NOTATION KEYS

Key	Description
a_{ij}	autonomic value contribution for individual metric j of functionality i
k_i	autonomic value contribution for individual functionality i
LoA	total level of autonomicity measure for all f_i & m_{ij}
M_i	number of metrics for functionality i
M_c, M_h, M_o, & M_p	number of metrics for each of the self-* functionalities respectively
m_{ij}	individual metric j for functionality i
n	number of functionalities
n_i	individual functionalities
r	possible combinations of functionalities
R_i	rank of a functionality i in the autonomic composition of a system
v_i	weighting for functionality i
w_{ij}	weighting for metric j of functionality i
c_i, h_i, o_i and p_i	autonomic metric contributions of the functionalities for a CHOP-based system
All indices (i and j) begin at 1	

A. Preliminary Work: A Specific Case Approach

To make progress in this approach, a preliminary effort is set out in [1]. This initial effort works perfectly well in cases where functionalities are orthogonal and for specific systems of limited (known) number of functionalities. Now, following on from equation (4) and taking a specific system in isolation, for example, (say a system with only four functionalities, e.g., the CHOP), this will give 16 possible combinations as shown in Figure 2, although the 16th combination is a special case which implies the system demonstrates no autonomic functionality and thus it is not considered further. The CHOP

functionalities may not all be of equal importance to a particular application domain hence we enumerate the possible combinations of functionalities, for reference. Combination 2 means that only two functionalities are of importance to the system's domain –so for example {C, H, not O, not P} is a specific combination representing a system with enumeration 4 in Figure 2.

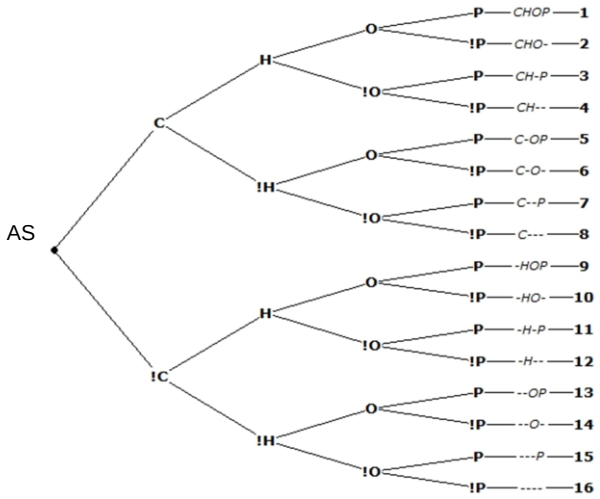


Figure 2: Combination of autonomic functionalities (for n = 4).

Figure 2 implies that, in terms of autonomic functionality composition, a system deemed autonomic (within the self-CHOP boundary) can be defined (or described) in one of n^2-1 ways. The remaining combination (enumerated 16 in Figure 2) represents a non-autonomic system, as it exhibits none of the autonomic functionalities. If we define autonomic metrics for each of the functionalities, then the sum of the autonomicity in each of the constituent functionalities for a particular AS gives the system's *LoA* (5). For example, the *LoA* of a system represented by line 9 in Figure 2 will be the summation of the autonomic metrics defining the self-healing, self-optimising and self-protecting functionalities.

$$LoA = \sum_{i=1}^{Mc}[c_i] + \sum_{i=1}^{Mh}[h_i] + \sum_{i=1}^{Mo}[o_i] + \sum_{i=1}^{Mp}[p_i] \quad (5)$$

Subscripted *M* is the number of identified metrics for the respective functionalities. c_i, h_i, o_i and p_i are the autonomic metric contributions of the functionalities. These can be composed of functions of different measures but as explained in Section IV(C), they are normalised to yield autonomic values. For more details regarding the preliminary work and the specific case approach of the proposed measure, see [1].

B. Measuring LoA: A Generic Case Approach

Having looked at specific (of known number of functionalities) case instance of the proposed approach, we seek, in this subsection, to establish a generic case in which this approach is suited for application across different scenario instances. Now, extending the approach and making it more generic, weighting is introduced. Functionalities are not necessarily orthogonal –i.e. a single behaviour could enhance the contribution of more than one metric and this could be across more than one functionality. This is important because the measurement approach has to work in situations where the functionalities are and are not orthogonal. In cases of non-orthogonality, the weighting is applied to tune sensitivity of contributing behaviours.

For flexibility of applying the technique across different application instances, *LoA* is normalised to a value in the range 0 to 1. It also follows that all autonomic value contributions and weighting are normalised within the same interval range:

$$0.0 \le LoA \le 1.0, \;\; 0.0 \le a_{ij} \le 1.0$$
$$0.0 \le v_i \le 1.0, \;\; 0.0 \le w_{ij} \le 1.0 \quad (6)$$

Normalisation of the individual components of the formulae is important to enable comparison of different systems with different implementations, and also to address non-linearity aspects. The way we measure the system should not on its own change the outcome –for example, higher number of metrics should not result in higher *LoA* value and as well does not translate to being 'more autonomic'. So in all cases, and for normalisation purposes, the following rules must apply:

$$\sum_{j=1}^{M_i} w_{ij} = 1.0, \qquad \sum_{i=1}^{n} v_i = 1.0\,, \qquad \sum_{j=1}^{M_i} a_{ij} = 1.0 \quad (7)$$

The metric weighting (w_{ij}) and metric autonomic value contribution (a_{ij}) are both with reference to individual functionalities and so are bound to the number of metrics for those functionalities (M_i). However, the functionality weighting (v_i) is with reference to the system itself and so is bound to the total number of functionalities (n) .This explains why the total individual autonomic value contribution ($\sum k_i$) can go up to n –see equation (9). If we ignore, for now, all indices and have a top level view of the proposed *LoA* calculation, for a single functionality, then:

$$k = (a \times w) \le 1.0 \quad (8)$$

$$\sum k \le n \sum \quad (9)$$

$$LoA = \sum(k \times v)$$
$$\rightarrow \sum[(a \times w) \times v] \;\; \forall a, w, v \le 1.0 \quad (10)$$

Decomposing (9) and (10) above, and for total autonomic value contribution of all functionalities n_i:

$$k_i = \sum_{j=1}^{M_i}\left(a_{ij} \times w_{ij}\right) \quad \forall n_i and\, m_{ij} \quad (11)$$

And applying the functionality weighting to the individual autonomic value contribution (k_i), we have:

$$k_i = v_i \times \left(\sum_{j=1}^{M_i} (a_{ij} \times w_{ij}) \right) \quad \forall n_i and\ m_{ij} \qquad (12)$$

LoA is then given by summing equation (12) for all values of n_i and m_{ij}:

$$LoA = \sum_{i=1}^{n} \left(v_i \times \left(\sum_{j=1}^{M_i} (a_{ij} \times w_{ij}) \right) \right) \qquad (13)$$

In the case of orthogonality or where weighting is not required, level of autonomicity is given by the basic expression:

$$LoA = \sum_{i=1}^{n} \sum_{j=1}^{M_i} (a_{ij}) \qquad (14)$$

This is equivalent to equation (5). Procedure 1 is a basic algorithm of the implementation of the proposed measure of autonomicity.

Procedure 1: Algorithm for implementing LoA

```
Input (main) variables: n and Mi
i = 1, 2, ..., n and j = 1, 2, ..., Mi
    if at n1, M1 = 3, then j = 1, 2, 3
  k1 = (w11 × a11) + (w12 × a12) + (w13 × a13)
k(1) = 0 //initialising k array
for i = 1 to n
 for j = 1 to M(i)
  sum(j) = w(ij) × a(ij)
  k(i) = k(i) + sum(j)
 next j
next i
 LoA = (k1 × v1) + ... + (kn × vn)
```

Note that the proposed approach is a 2-dimensional definition. That is, it supports only two levels of description, e.g., a system on one hand and its functionalities or characteristics on the other hand. A bit of tweaking and adaptation is required to support higher dimensional definitions e.g., a system, its functionalities or characteristics, sub-functionalities or sub-characteristics, etc.

C. Normalisation and Scaling of Autonomic Metrics Dimensions

There is still a point though that needs to be addressed. When computing for *LoA*, we are normalising values that are products of aggregated metric values of different units and dimensions. Depending on the application domain, metrics can be scalar (of different measures) or non-scalar values (e.g., observing a capability, Boolean based decisions, etc.). So, despite what measure or form these metrics take, there needs to be a way of scaling the metric values (of all contributing metrics) to a centric unit of *autonomic metric contribution* within a certain normalised range. But, because the range of values and metrics can vary significantly, each choice of how these are scaled can influence very differently the final *LoA*. A possible solution is to define scaling factors for all contributing metrics within a normalised range (of [0, 1] in our case). In this way, the metrics' values (irrespective of units of measure) are normalised into real numbers that are summed to give *LoA*. One challenge here, though, is defining the scaling factors. We identify two simple methods for normalisation: 1) By ranking values according to *high*, *medium*, and *low*. The meaning of this ranking is metric-dependent and is based on a defined margin. For example, if a maximum expected value is 6, a value of 0-2 will be ranked *low*, while 3-4 will be ranked *medium* and 5-6 ranked *high*. A medium value would contribute fifty percent of the metric's autonomic value contribution in the range of [0, 1] (recall that $0.0 \leq a_{ij} \leq 1.0$ from equation (6)), while the two extremes would contribute zero and hundred percents –these may differ depending on choice of usage. This can be used for scalar metrics like the *interoperability ratio* and *reaction time* metrics discussed in Section III. 2) By having a Boolean kind of contribution where two values can suggest two extremes – either affirming a capability or not. For example, if a '*True*' outcome affirms a capability then it contributes hundred percent of the autonomic value contribution, while a '*False*' outcome contributes zero. Another example in this category is where an instance of an event either does or does not confirm a capability (e.g., the *stability* metric for self-optimising). Other specific methods, like the *Mahalanobis Distance* [22] discussed and used in [23], have been proposed. In scaling the different dimensions of distances between points (measured in different distance measurement units), the authours of [23] use a simplified form of the *Mahalanobis Distance*, where for each dimension, they compute the standard deviation over all available values and then express the components of the distances between points as multiples of the standard deviation for each component.

In the end, anyone can choose any form of scaling and normalisation as long as it is uniformly used across board for all systems to be evaluated and all values are within the range [0, 1] as explained in equations (6) and (7).

D. Measuring LoA: Comparison of Approaches

Assessing autonomic systems and being able to analyse and compare diverse systems of different degrees is an open research challenge that needs significant attention. There have been several attempts to develop a way of measuring autonomicity but unfortunately a universal solution has not been found. [8] shows that up to this point, there is one main approach to measuring the extent of autonomicity of autonomic systems (the *scale-based* approach which is explained in Section II), and a number of variations of this

have been explored. The fundamental purpose of this approach is to reflect the level of involvement in decision-making between the system and the human user. The major variations of the *scale-based* approach are Clough [4], Sheridan [9], and Ryan *et al.* [2]. Clough's 10-level scale is a result of developing national intelligent autonomous UAV metrics for the Department of Defence (DoD). Though it is tied to UAVs, its use of metrics to measure the level of autonomy of UAVs makes it stand out. The levels in-between the scale are populated by defining metrics for UAVs. This is good because using metrics that define functionalities gives a clearer understanding of the systems. Yet there is no normalised single point of reference that can be used in comparing two systems using this approach. Sheridan's 10-level scale measures two aspects; decision making (levels 2-4) and decision execution (levels 5-9). Ultimately, Sheridan focuses on human-machine relations (and human supervisory control) and not necessarily on the level of autonomicity of systems. Ryan *et al* extended Sheridan's concept and developed an 8-level scale that determines the level of autonomy *needed* in designing autonomic systems; although their work cannot actually be said to offer a way of *measuring* autonomic systems' level of autonomicity.

None of these is sufficiently sophisticated in measuring *LoA*. The technique we propose here is more sophisticated in a number of ways: it is the only technique that ties down *LoA* to a numeric value; it takes into account individual weights; it is flexible in the sense that it can take any number of degrees (functionalities), and the fact that the numeric value is scaled always to a normalised value (to cater for comparisons between systems with different numbers of dimensions of autonomicity and different numbers of metrics for measuring the extent of functionality achieved in each dimension). Normalisation gives you the power to compare two different systems no matter the number of individual metrics.

E. Evaluating Autonomic Systems

Evaluating Autonomic Systems using equation (5) or equation (13) gives their separate *LoA* values –which are aggregated values. This, however, does not give a fine-grained picture of the systems' performances in terms of individual functionalities. Systems are classified according to their implementation combinations (*r*). This is in terms of what self-* functionalities are required or demonstrated in their specific application domains. One thing remains to be clarified at this point –'how do we rank each functionality in the autonomic composition of a system?' This can be in terms of importance or extent of functionality provided. We focus on the later –the extent of functionality provided as against what is needed. Take for instance, if two systems are of the same combination we may wish to know which of them provides a greater degree of say self-healing or self-protection in any application domain. To address this, we adapt a function that measures agent's decision-making power in a multi-agent autonomic system defined in [11]. The rank of a functionality R_i in the autonomic composition of a system is defined by the ratio of its autonomic contribution (k_i or a_{ij}) to the total autonomic contribution of all metrics defining the composite functionalities of that system:

$$R_i = \frac{k_i}{LoA} \tag{15}$$

This applies where weighting is considered. If weighting is not considered, R_i is given by equation (16):

$$R_i = \frac{\sum_{j=1}^{M_i} a_{ij}}{LoA} \tag{16}$$

where (k_i or a_{ij}) is the autonomic contribution of the considered functionality which could be the summation of c_i, h_i, o_i or p_i in equation (5) or the calculation of k_i in equation (11) or the summation of a_{ij} (e.g., the case in equation (14)). With equations (15) and (16), any composite functionality can be ranked in terms of their autonomic contribution.

V. AUTONOMIC SYSTEMS EVALUATION CASE STUDY

In this section, two example cases that cover the specific and generic case approaches (explained in Section IV) are examined. The first is based on *Dynamic Qualitative Sensor Selection System* (DQSSS) application scenario (see [19] for full details of the DQSS system). This is used to demonstrate a case where functionalities are assumed to be orthogonal and for specific systems of fixed number of autonomic functionalities. This is consistent with the preliminary work in [1] and suites the proponents of the view that autonomic systems are only defined by the generally accepted and core functionalities of the self-CHOP.

The second case example deploys one of the current technology innovations –Autonomic Marketing. This is used to demonstrate a generic case instance where functionalities are not necessarily orthogonal and where systems are defined by *n* number of autonomic functionalities. For more details on the autonomic marketing system scenario see [24] and [25].

For each case example, three systems (or autonomic managers) are examined. When comparing these systems, it is important to look closely at the performances of individual autonomic functionality to give clearer understanding of the calculated *LoA*.

A. Dynamic Qualitative Sensor Selection Case Example

In this example, autonomic functionalities are limited to the original, and generally accepted four self-CHOP functionalities, supposing that any autonomic system is defined by them. This is representative of many real-world systems of known (fixed) functionalities or characteristics. The DQSSS case study is based on work in [19]. The goal of DQSSS is to *dynamically select a sensor* (amongst many) *based on continuously variable qualitative characteristics* (e.g., signal quality and noise levels). This is typical of an application that accesses several sensors generating raw data from monitoring a particular context; these could be physical attributes of a system or perhaps information feeds from a

service (e.g. financial data). In such applications, it is expected that a DQSSS would generate and differentiate signal characteristics and trends, choose the best signal and without compromising stability, be continuous, unsupervised, dynamic, and detect and react if a sensor goes down. Autonomic metrics are drawn from these characteristics. By definition, self-configuration, self-optimization and self-healing are of importance to this system (i.e., *r=3* and also *n* is fixed at 4). The DQSSS in [19] is presented in three progressive stages which we refer here to as systems A, B and C. All three systems are able to differentiate sensors by their signal characteristics such as noise level and spikes. These are then combined in a *utility function* to determine the better quality sensor. Systems B and C are able to generate trends in signal quality using *trend analysis* logic. Only system C ensures stability (avoiding unhealthy oscillation in sensor selection) by implementing *dead zone* logic, while none of the systems has a way of detecting a failed sensor.

TABLE II: REPRESENTATION OF THE DQSSS [19]

Characteristics (metrics)	**Contributing CHOP**	**Sys A**	**Sys B**	**Sys C**
Continuous	C	√	√	√
Unsupervised	C	√	√	√
Trends examination	O	-	√	√
Stability	O	-	-	√
Dynamic (logic switching)	O	-	-	√
Signal characteristics	C	√	√	√
Signal differentiation	C	√	√	√
Failure sensitivity (sensors)	H	-	-	-
Robust (fault tolerance)	H	-	-	√

In keeping with the normalisation of values as contained in (6) and (7), the maximum achievable *LoA* becomes '1' implying that each CHOP contributes an autonomic value in the range ($0 \leq a_{ij} \leq \frac{1}{M_i}$) spread across its metrics. Normalising the identified metrics in Table II (the numbers of metrics in each combination are: $M_1 = 4$ (for self-configuring C), $M_2 = 2$ (for self-healing H), and $M_3 = 3$ (for self-optimising O)) in the autonomic value range ($0 \leq a_{ij} \leq \frac{1}{M_i}$) and applying equation (5) or (14) gives the result in Table III. Equation (17) is an expression of how each instance of the metrics contribution is calculated.

$$a_{ij} = \left(\frac{1}{n}\right) \times \left(\frac{1}{M_i}\right) \tag{17}$$

Figure 3 is a radar chart analysis of systems A, B and C in terms of their separate autonomic functionality composition. Recall that only three functionalities (CHO-) are of importance here which explains why self-protection P has no value. Based on the *LoA* achievements of the three systems A, B and C as shown in Table III (0.25, 0.33 and 0.63 respectively), it means that in a dynamic sensor selection application domain (as defined), system C can be depended upon to carry out the task with a higher confidence level and lower risk factor compare to systems B and A.

One powerful aspect of our proposal, particularly the specific case approach with fixed number of functionalities, is that it offers the flexibility of qualitatively interpreting *LoA* results using any *scale-based* approach. This is done by applying the upper bound of the chosen scale to equation (17) as in equation (18) and then interpreting the results within the levels of the scale.

$$a_{ij} = \left(\frac{max}{n}\right) \times \left(\frac{1}{M_i}\right) \tag{18}$$

Where *max* is upper bound of the scale used.

Applying Ryan *et al* level of autonomy assessment scale [2] which, as explained in Related Work section, is an 8-level autonomy assessment tool (used for either identifying (qualitatively) the level of autonomy of an existing system or for proposing an appropriate level of autonomy during the design of a new system), *max* becomes 8. So, in computing (18) with *max* = 8, system A falls within level 2 of the scale which points to a situation where '*computer shadows human*' in the self-management process. This indicates that system A only has a narrow envelope of environmental conditions in which it is both autonomic and returns satisfactory behaviour. System B tends toward level 3 on the scale which is '*human shadows computer*' which translates into a wider operational envelope, but once the limits of that envelope are reached human input is needed in the form of retuning, or manual override in the case of oscillation, which for example system C can deal with autonomically. System C falls within level 5, which points to '*collaboration with reduced human intervention*'. This indicates that C is sufficiently sophisticated to operate autonomically and yield satisfactory results under almost all perceivable operating circumstances.

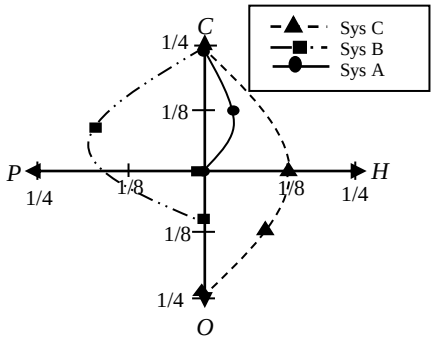


Figure 3: LoA representation of systems A, B & C in the CHOP domains.

TABLE III: ANALYSIS RESULT

	Sys A	Sys B	Sys C
C	0.25	0.25	0.25
H	0.00	0.00	0.13
O	0.00	0.08	0.25
P	0.00	0.00	0.00
LoA	**0.25**	**0.33**	**0.63**

Employing (16) to rank the functionalities and taking just self-configuration for example, we find that in system A,

self-configuration contributes 100% of its autonomic achievement, while in systems B and C the contribution is 75% and 40% respectively. This provides another analytic spectrum that gives a clearer understanding of the composition of the calculated *LoA*.

The benefit of analyzing Autonomic Systems in terms of their extent of autonomicity not only offers a path to Autonomic Systems' certification as stated earlier, it also offers a way of comparing these systems, and also facilitates a proper description of these systems to users.

B. Autonomic Marketing Case Example

In this example, we consider a specific aspect of autonomic marketing system based on the work and experiment presented in [25]. As there are yet no standardised (or defined) lists of autonomic metrics, at least for the case example system, we draw autonomic metrics and functionalities for the purposes of this case example from the specified goal of the system as detailed in [25]. So, metrics are drawn based on (or limited to) what is explained in the experiment and not on what autonomic marketing systems, generally, should have as metrics. In the end, the interest here is to show how the proposed *LoA* measurement approach can be applied to systems.

The particular case example autonomic marketing system studied here is that of targeted television advertising during a live sports competition airing. A company is interested in running an adaptable marketing campaign on television with different adverts (of different products appealing to audiences of different demographics) to be aired at different times during a live match between two teams. There are three adverts (Ad1, Ad2 and Ad3) to be run and the choice of an ad will be influenced by, amongst other things, viewer demographics, time of ad (local time, time in game, e.g., half time, TV peak/off-peak time, etc.), length of ad (time constraint), cost of ad, who is winning in the game, etc. This is a typical example of a system with many dimensions of freedom and very wide behaviour space. The behaviour space is divided into four zones along two dimensions of freedom (*Mood* and *CostImplication*) as shown in Figure 4.

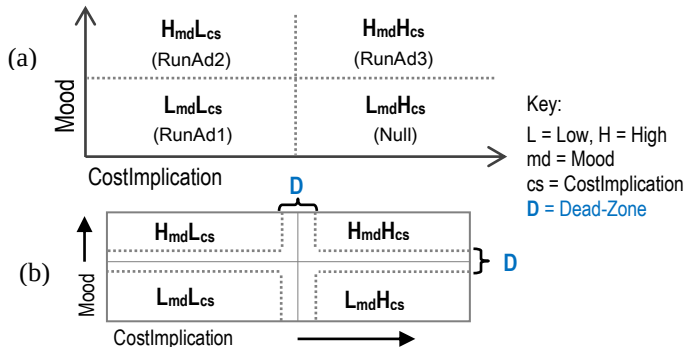


Figure 4: System behaviour space in 2 dimensions of freedom [25]

The two dimensions of freedom, which are influenced by several contextual variables, represent a collation of all possible decision influencers. Each action (ad) is thus activated only in its allocated zone following specified policy in order to achieve the system's goal defined by a set of rules (Figure 5). Following the set policy, the autonomic manager, at every *decision* instance (of a sample collection) decides on which ad to run. The optimisation of the system is in terms of achieving balance between efficient just-in-time target-marketing decision and cost effectiveness (savings maximisation) while maintaining improved trustability, stability and dependability in the process.

1. Extract external variables (decision parameters) at defined time interval and decide on action
2. Send trap msg and change action if (*condition omitted*) otherwise retain previous action
3. If current action is same as previous action, do not send trap and do not change action

==========Measure of Success==========

4. Cost of action change (total ad run) must fall within budget
5. Rate of change should be considerably reasonable
6. Maximum of one ad change within the first five sample collections and subsequently maximum of two in any three sample instances
7. Turnover should justify cost

Figure 5: Excerpt of rules defining system goal [25].

Three autonomic managers, based on three different levels of autonomic architectures, are designed to implement this system. In this example, we evaluate these three managers (full detail of experiment is available in [25]). Basically, the first manager, AC (AutonomicControlling at its core), is concerned with making decisions within the boundaries of the rules while the second, VC (ValidationCheck at its core) goes beyond decision making to validate decisions for conformity with the rules. The third manager, DC (DependabilityCheck at its core) verifies that the measure of success is achieved. DC also improves reliability by instilling stability in the system. This is done by introducing dead-zone boundaries (Figure 4b) within which, no action is taken (avoiding erratic and unnecessary changes) and implementing a TRC (Tolerance-Range-Check) to address rules in particular.

Based on the goal of the system, specified in the rules of Figure 5, we have drawn three functionalities (self-configuration, self-optimisation and self-stability) and six metrics. Table IV shows the metrics and contributing rules. The simulation was run for a total duration of 50 sample collection instances. This means that for the duration of the simulation, external variables (that influence decisions for ads) were fed to the autonomic managers fifty times.

TABLE IV: AUTONOMIC METRICS FOR SYSTEMS AC, VC & DC

Metrics	Description	Contributing rule
# of ad change (x)	Number of times ads changed	Rule 2
# of ad run (y)	Number of ads that were run	Rules 3 and 4
# of decision (d)	Number of decision instances	d is a constant
Rate of ad change (T_x = x/50)	Rate at which ads were changing	Rules 2 and 5
Rate of ad run (T_y = y/50)	Rate at which ads were run	Rules 4 and 7
Decision ad change ratio (d_x = x/d)	Number of decision ad change relation	Rule 3
Stability (s)	Number of times TRC bounds were breached	Rules 5 and 6

Recall that functionalities are not always orthogonal as in this example some metrics contribute to more than one functionality (see Table V). The rate of the influence of the metrics on the functionalities is tuned by applying weighting. Table V shows the autonomic value contributions of all metrics and the corresponding functionalities. Weights are discretely allocated to reflect relevance and importance of functionalities (based on the goal of the system). Metric values are averages of 10 different simulation runs of the 50 sample collection instances (see Appendix A for more details). In this example, we assume all metrics to be of equal weight within their respective functionalities.

TABLE V: DISTRIBUTION OF METRIC VALUES

Functionality (n_i)	Weight (v_i)	Metric (M_i)	Metric weight (w_{ij})			Metric contribution (a_{ij})		
			AC	VC	DC	AC	VC	DC
Self-configuration	0.20	x	0.50	0.50	0.50	0.000	2.000	4.800
		T_x	0.50	0.50	0.50	0.768	0.808	0.864
Self-optimisation	0.40	y	0.20	0.20	0.20	2.600	3.900	5.800
		T_y	0.30	0.30	0.30	11.42	11.45	11.49
		s	0.30	0.30	0.30	9.600	10.20	11.60
		d_x	0.20	0.20	0.20	0.000	0.171	0.414
Self-stability	0.40	T_x	0.25	0.25	0.25	0.768	0.808	0.864
		y	0.05	0.05	0.05	2.600	3.900	5.800
		T_y	0.20	0.20	0.20	11.42	11.45	11.49
		s	0.50	0.50	0.50	9.600	10.20	11.60

From the values of Table V, we can now calculate the *LoA* of all three systems (AC, VC, and DC). Because of space we will only show the calculations for that of system AC and then use the LoA Calculator (see Section V C) to calculate the rest.

$n = 3$

For n_1: $M_1 = 2$, $v_1 = 0.20$, $w_{11} = 0.50$, $w_{12} = 0.50$, $a_{11} = 0.00$, and $a_{12} = 0.768$

For n_2: $M_2 = 4$, $v_2 = 0.40$, $w_{21} = 0.25$, $w_{22} = 0.25$, $w_{23} = 0.25$, $w_{24} = 0.25$, $a_{21} = 2.600$, $a_{22} = 11.42$, $a_{23} = 9.600$, and $a_{24} = 0.000$

For n_3: $M_3 = 4$, $v_3 = 0.40$, $w_{31} = 0.25$, $w_{32} = 0.25$, $w_{33} = 0.25$, $w_{34} = 0.25$, $a_{31} = 0.768$, $a_{32} = 2.600$, $a_{33} = 11.42$, and $a_{34} = 9.600$

$$k_1 = (a_{11} \times w_{11}) + (a_{12} \times w_{12}) = (0.00 \times 0.50) + (0.768 \times 0.50)$$
$$= (0.00) + (0.384)$$
$$= 0.384$$

$$k_2 = (a_{21} \times w_{21}) + (a_{22} \times w_{22}) + (a_{23} \times w_{23}) + (a_{24} \times w_{24})$$
$$= (2.600 \times 0.25) + (11.42 \times 0.25) + (9.60 \times 0.25) + (0.0 \times 0.25)$$
$$= (0.65) + (2.80) + (2.40) + (0.00)$$
$$= 5.850$$

$$k_3 = (a_{31} \times w_{31}) + (a_{32} \times w_{32}) + (a_{33} \times w_{33}) + (a_{34} \times w_{34}) =$$
$$(0.768 \times 0.25) + (2.600 \times 0.25) + (11.42 \times 0.25) + (9.60 \times 0.25)$$
$$= (0.192) + (0.650) + (2.855) + (2.400)$$
$$= 6.097$$

Applying equation (13):

$$\text{LoA} = (k_1 \times v_1) + (k_2 \times v_2) + (k_3 \times v_3)$$
$$= (0.384 \times 0.20) + (5.850 \times 0.40) + (6.097 \times 0.40)$$
$$= 0.0768 + 2.340 + 2.439$$
$$= 4.8558$$

Figure 6 is a snapshot of the *LoA* Calculator's result console showing the *LoA* results of systems VC and DC. Recall that the choice of scaling and normalisation used can influence very differently the final *LoA*. In making a choice, the nature of system and metrics need to be considered. In this example, we can choose to normalise the individual sub-columns of the a_{ij} column within the range ($0 \leq a_{ij} \geq 1$), but this will negatively affect the values across all three systems and make it difficult for *LoA* calculations. Also we cannot normalise within the column across rows as that will overshoot the normalisation range within the sub-columns. So, we calculate with raw values and then normalise the final *LoA* values. How the metric values in Table V are scaled is shown in Appendix A.

From Figure 6:

LoA for system VC = 5.4887
LoA for system DC = 6.4722

The calculated *LoA* for system AC = 4.8558

Then, normalising all three values within the normalisation range of ($0 \leq LoA \geq 1$) using expression (19):

$$AC = \frac{AC}{(AC + VC + DC)}, VC = \frac{VC}{(AC + VC + DC)}, DC = \frac{DC}{(AC + VC + DC)} \quad (19)$$

For system AC

$$LoA = \frac{4.8558}{16.8167} = \mathbf{0.2887}$$

For system VC

$$LoA = \frac{5.4887}{16.8167} = \mathbf{0.3263}$$

For system DC

$$LoA = \frac{6.4722}{16.8167} = \mathbf{0.3849}$$

The results clearly indicate the superiority of the systems from DC down to AC in terms of level of autonomicity (based on the criteria set as system goal). What this means is that, in terms of the criteria specified as the goal of the system, the autonomic manager of system DC is more autonomic than the others followed by system VC. The margins reflect, almost with the same magnitude, the performances of the systems as results show in [25].

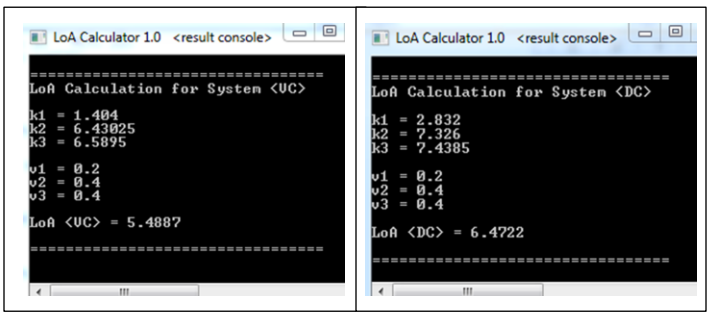


Figure 6: LoA calculate console result for systems VC and DC

Recall that the optimisation of the system in question is in terms of achieving balance between efficient just-in-time target-marketing decision and cost effectiveness (savings maximisation) while maintaining improved trustability, stability and dependability in the process. From experimented results, system DC shows significant gain in stability and cost savings. It also smoothened the high fluctuation rate (high adaptability frequency) experienced by other systems and in general, reduces the average ad change ratio of about one change in three sample collections (1:3) to one change in ten sample collections (1:10), representing an overall gain of about 31.25% in terms of stability and cost efficiency.

C. LoA Calculator

The LoA Calculator is an application that helps in calculating system's level of autonomicity. The application is developed using C# and can be used in calculating the LoA of any system at any level of complexity. The application can be used for both the generic and specific case approaches. For the specific case approach (that may require no weights), the user enters the value '1' (one) in the place of all weights and that will automatically cancel the weighting effects. Basically, all variables and values used are user-defined and are fed into the application for LoA computation. There are two formats for the application. The basic format is for simple systems of few variables and provides a dialogue interface (a form) for a user to enter system variables. This is suitable when there are only a few data to be fed into the application and can be done through the keyboard. Figure 7 is a snapshot of the dialogue interface. The other format is more complex and is used for complex systems (of multiple data). The complex format feeds data into the application using comma separated text file (csv file). The user specifies raw data in a CSV file template and provides the file path to the application during run-time. Both formats work on the same principle (the core is based on equation (13)) and can be used interchangeably but it is more tedious using the basic format for complex systems.

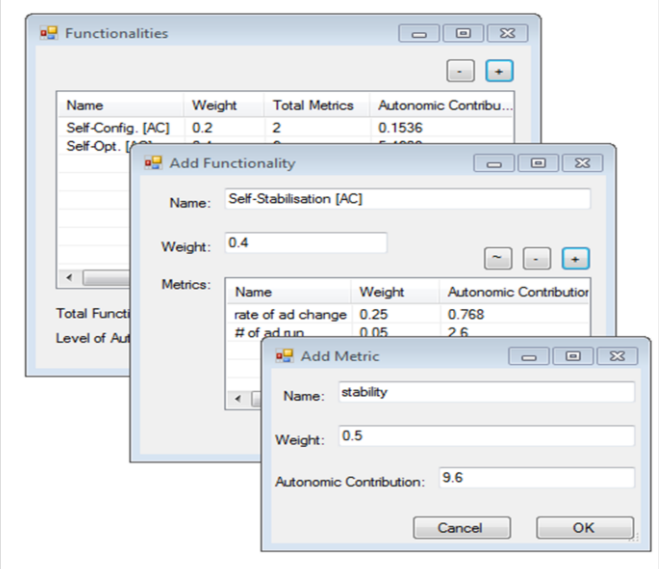


Figure 7: LoA Calculator basic format dialogue interface

In the current (first) version of the application, only variables and values for one system can be fed into the system at any one time. Subsequent versions will be able to take values for more than one system at one instance and evaluate the systems in terms of their separate LoAs. The application is available for download at [26].

VI. CONCLUSION AND FUTURE WORK

A system is better defined by its capabilities and so measuring the LoA of Autonomic Systems without a reference to autonomic functionalities would be inaccurate. We have proposed a functionality-based LoA measurement. In our proposal, a typical AS is defined by some core autonomic functionalities and LoA is measured with respect to these functionalities. Each functionality is defined by a set of metrics. The metrics values are normalised and aggregated to give the autonomic contribution of each functionality which are then combined to yield a LoA value for an AS. Our proposed approach is in two forms; the specific case approach and the generic case approach. The specific case approach works perfectly well in cases where functionalities are orthogonal and for specific systems of limited (fixed) number of functionalities. We have shown how this approach can adapt any scale-based approach to enable a qualitative understanding of the quantitative LoA measure proposed here. The generic case approach is used to demonstrate a generic case instance where functionalities are not necessarily orthogonal and where systems are defined by *n* number of autonomic functionalities. We have also shown how systems can further be evaluated to give a fine-grained picture of the systems' performances in terms of individual functionalities looking at the ratio of autonomic contributions of their separate functionalities. In this, we found that only systems within the same implementation combination can be compared. We have carried out two case study examples (for the specific and generic case approaches) to demonstrate the usage and applicability of our proposed LoA measure. There are several other research works trying to develop a way of measuring autonomicity but have not succeeded. Some approaches have been proposed but none of these is sufficiently sophisticated in measuring LoA. Our technique here is more sophisticated in a number of ways: the fact that it is the only one that ties down LoA to a numeric value, the fact that it takes into account individual weights, it is flexible in the sense that it can take any number of degrees (properties), and the fact that numeric values are scaled always to a normalised value – (which otherwise gives the wrong impression that more metrics mean more autonomicity. Normalisation gives you the power to compare two different systems no matter the number of individual metrics).

The standardization of a technique for the measurement of LoA will bring many quality-related benefits which include being able to compare alternative configurations of autonomic systems, and even to be able to compare alternate systems themselves and approaches to building autonomic systems, in terms of the LoA they offer. This in turn has the potential to improve the consistency of the entire lifecycle of Autonomic

Systems and in particular links across the requirements analysis, design and acceptance testing stages.

One research challenge is the study and standardisation of autonomic metrics for different autonomic systems. The metrics definitions can be grouped or modularised (e.g., the standardised categorisation of UMS in [20]). So, as future work, we are looking at standardised ways of properly defining and generating autonomic metrics to strengthen our framework. This is a key component towards our wider research which focuses on the challenge of validating autonomic computing systems to achieve trustworthiness. We will be developing more sophisticated versions of the LoA Calculator.

REFERENCES

[1] Eze T., Anthony R., Walshaw C. and Soper A. *A Technique for Measuring the Level of Autonomicity of Self-managing Systems.* In proceedings of The 8th International Conference on Autonomic and Autonomous Systems (ICAS 2012), St. Maarten, The Netherlands Antilles, March 2012

[2] Proud R., Hart J., and Mrozinski R. *Methods for Determining the Level of Autonomy to Design into a Human Spaceflight Vehicle: A Function Specific Approach*. Report date September 2003 http://handle.dtic.mil/100.2/ADA515467 accessed 04/09/2012

[3] Eze T., Anthony R., Walshaw C. and Soper A. *The Challenge of Validation for Autonomic and Self-Managing Systems.* In proceedings of The 7th International Conference on Autonomic and Autonomous Systems (ICAS), May 22-27, 2011 – Venice/Mestre, Italy

[4] Clough B. *Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway?* In Proceedings of PerMis Workshop, pp 1–7. NIST, Gaithersburg, MD, 2002.

[5] IBM Autonomic Computing White Paper. *An architectural blueprint for autonomic computing*. 3rd edition, June 2005

[6] ISO/IEC 9126-1:2001(E). Software engineering — Product quality — Part 1: Quality model

[7] Honeycutt G. *How Much Do we Trust Autonomous Systems?* Unmanned Systems -2008

[8] Eze T., Anthony R., Walshaw C. and Soper A. *Autonomic Computing in the First Decade: Trends and Direction*. In proceedings of The 8th International Conference on Autonomic and Autonomous Systems (ICAS), St. Maarten, The Netherlands Antilles, March 2012

[9] Sheridan T. *Telerobotics, Automation, and Human Supervisory Control*. The MIT Press. Cambridge, MA, USA 1992. ISBN:0-262-19316-7

[10] Huebscher M. and McCann J. *A survey of autonomic computing—degrees, models, and applications*. ACM Computer Survey, 40, 3, Article 7 (August 2008)

[11] Barber, K. and Martin, C. *Agent Autonomy: Specification, Measurement, and Dynamic Adjustment*. In Proceedings of the Autonomy Control Software Workshop at Autonomous Agents 1999 (Agents'99), 8-15. Seattle

[12] Alonso F., Fuertes J., Martínez L., and Soza H. *Towards a Set of Measures for Evaluating Software Agent Autonomy*. In proceedings of 8th Mexican Int'l Conference on Artificial Intelligence (MICAI), 2009

[13] Kephart J., and Chess D. *The Vision of Autonomic Computing*. Computer, IEEE, Vol 36, Issue 1, 2003, pp 41-50

[14] McCann J. and Huebscher M. *Evaluation issues in Autonomic Computing.* In proceedings of Grid and Corporative Computing (GCC) Workshop, LNCS 3252, pp. 597-608, Springer-V erlag, Birlin Heidelber, 2004

[15] Bantz D., Bisdikian C., Challener D., Karidis J., Mastrianni S., Mohindra A., Shea D., and Vanover M. *Autonomic Personal Computing*. IBM Systems Journal, Vol 42, No 1, 2003

[16] Mark B., Sheng M., Guy L., Laurent M., Mark W., Jon C., and Peter S. *Quickly Finding Known Software Problems via Automated Symptom Matching*, The 2nd International Conference on Autonomic Computing (ICAC), 2005, Seattle, USA

[17] Tianfield H. *Multi-agent Based Autonomic Architecture for Network Management*. In Proc. IEEE International Conference on Industrial Informatics, pp. 462–469, 2003

[18] Truszkowski W., Hallock L., Rouff C., Karlin J., Rash J., Hinchey M., and Sterritt R. *Autonomous and Autonomic Systems*. Springer, 2009

[19] Anthony R. *Policy-based autonomic computing with integral support for self-stabilisation*, Int. Journal of Autonomic Computing, Vol. 1, No. 1, pp.1–33. 2009

[20] Huang H., Albus J., Messina E., Wade R., and English W. *Specifying Autonomy Levels for Unmanned Systems: Interim Report*, SPIE Defense and Security Symposium 2004, Conference 5422, Orlando, Florida, April 2004.

[21] Huang H., Pavek K., Albus J., and Messina E. *Autonomy Levels for Unmanned Systems (ALFUS) Framework: An Update*. In proceedings of SPIE Defense and Security Symposium, Orlando, Florida. 2005

[22] Online article. *Mahalanobis Distance*, available via http://classifion.sicyon.com/References/M_distance.pdf viewed 10/09/2012

[23] Huebscher M. and McCann J. *An adaptive middleware framework for context-aware applications*, Springer Volume 10, Issue 1, pp 12-20, February 2006

[24] Adams C., Anthony R., Powley W., Bell D., White C., and Wu C. *Towards Autonomic Marketing*, The 8th International Conference on Autonomic and Autonomous Systems (ICAS), pp. 28-31, St. Maarten 2012.

[25] Eze T., Anthony R., Walshaw C. and Soper A. *A New Architecture for Trustworthy Autonomic Systems*. In proceedings of The 4th International Conference on Emerging Network Intelligence (EMERGING), Barcelona, Spain, 2012

[26] LoA Calculator Downloads via http://thaddeus-eze.com. Accessed 19/12/2012

[27] Parasuraman R., Sheridan T., and Wickens C. "A model for types and levels of human interaction with automation," IEEE Transactions on Systems, MAN, And Cyberneticsparta: Systems and Humans, vol. 30, pp. 286–297, MAY 2000

[28] Truszkowski W., Hallock L., Rouff C., Karlin J., Rash J., Hinchey M., and Sterritt R. Autonomous and Autonomic Systems. Springer, 2009

APPENDIX A

	# of ad change (x)			# of ad run (y)			stability			rate of ad change (x/50)			rate of ad run (y/50)			(d) # of	decision ad change ratio (x/d)		
Runs	AC	VC	DC	AC	VC	DC	AC	VC	DC	AC	VC	DC	AC	VC	DC	decisions	AC	VC	DC
1	12	12	7	7	7	5	2	2	0	0.24	0.24	0.14	0.14	0.14	0.1	12	1	1	0.5833333
2	8	6	3	7	5	3	3	2	0	0.16	0.12	0.06	0.14	0.1	0.06	8	1	0.75	0.375
3	15	11	8	12	8	6	4	2	0	0.3	0.22	0.16	0.24	0.16	0.12	15	1	0.73333	0.5333333
4	10	7	6	7	7	5	1	0	0	0.2	0.14	0.12	0.14	0.14	0.1	10	1	0.7	0.6
5	15	12	9	10	9	7	2	2	0	0.3	0.24	0.18	0.2	0.18	0.14	15	1	0.8	0.6
6	10	9	8	9	9	8	2	0	0	0.2	0.18	0.16	0.18	0.18	0.16	10	1	0.9	0.8
7	13	10	6	12	9	6	3	3	0	0.26	0.2	0.12	0.24	0.18	0.12	13	1	0.76923	0.4615385
8	11	11	7	7	7	5	3	3	0	0.22	0.22	0.14	0.14	0.14	0.1	11	1	1	0.6363636
9	11	9	7	10	9	7	0	0	0	0.22	0.18	0.14	0.2	0.18	0.14	11	1	0.81818	0.6363636
10	11	9	7	9	7	6	0	0	0	0.22	0.18	0.14	0.18	0.14	0.12	11	1	0.81818	0.6363636
Avg	**11.6**	**9.6**	**6.8**	**9**	**7.7**	**5.8**	**2**	**1.4**	**0**	**0.232**	**0.192**	**0.136**	**0.18**	**0.154**	**0.116**	**11.6**	1	**0.82889**	**0.5862296**

Table A1: Raw metric values as collected from experimental results

Table A1 shows the raw values of 10 different runs of the same simulation. It is not scientifically reliable to work with values of a single simulation.

Note that *Stability* measures the rate at which the tolerance range check (TRC) is breached. The column for DC is all zero because, in all 10 simulation runs DC did not breach the TRC bound.

Metric	AC	VC	DC		AC	VC	DC
# of ad change (x)	11.6	9.6	6.8	11.6-all i.e, max minus values	0	2	4.8
# of ad run (y)	9	7.7	5.8	11.6-all i.e, max minus values	2.6	3.9	5.8
stability	2	1.4	0	11.6-all i.e, max minus values	9.6	10.2	11.6
rate of ad change (x/50)	0.232	0.192	0.136	1-all i.e, max minus values	0.768	0.808	0.864
rate of ad run (y/50)	0.18	0.154	0.116	11.6-all i.e, max minus values	11.42	11.45	11.49
# of decisions (d)	11.6	11.6	11.6				
decision ad change ratio (x/d)	1	0.829	0.586	1-all i.e, max minus values	0	0.171	0.414

Table A2: Scaled metric values

Table A2 (left side) is a collation of the averages of the values of Table A1. On the right side of the table are the working values used. These are generated with reference to the possible maximum values of the individual metrics. For example, ads are changed or retained at every decision instance and since on the average, there are 11.6 decision instances, there are as much maximum possible number of ad change. So, the actual number of ad change is the difference between the number of possible ad changes and observed number of ad changes

The AC number of ad change metric is 11.6 (same as the number of decisions) indicating that AC changes ad at every decision instance. This tends to instability. Compare this to the value for DC and observe the difference.

Concepts and Mechanics for Educational Mini-Games

A Human-Centred Conceptual Design Approach involving Adolescent Learners and Domain Experts

Bieke Zaman, Yorick Poels, Nicky Sulmon, Jan-Henk Annema, Mathijs Verstraete, Dirk De Grooff
Centre for User Experience Research (CUO| Social Spaces, KU Leuven / iMinds)
Parkstraat 45 Bus 3605
3000 Leuven, Belgium
{bieke.zaman, yorick.poels, nicky.sulmon, janhenk.annema, mathijs.verstraete, dirk.degrooff}@soc.kuleuven.be

Frederik Cornillie, Piet Desmet
Interdisciplinary Research on Technology, Education & Communication (ITEC)
KU Leuven / iMinds Future Health Department
Etienne Sabbelaan 53
8500 Kortrijk, Belgium
{frederik.cornillie, piet.desmet}@kuleuven-kortrijk.be

***Abstract* — This article reports on two conceptual design sessions in which concepts for educational mini-games were generated through a human-centred approach. First, co-creation sessions were held with 14 adolescents between 14 and 16 years old in order to gain insight into their preferences for educational games for language learning. During these sessions, 11 game concepts were generated, revealing a classification of concepts for games oriented towards, on the one hand, formal language learning and, on the other hand, more informal communication with other players or in-game characters. Second, brainstorm sessions were organized with six domain experts in order to reveal which mechanics are most appropriate for the design of mini-games for a variety of educational programmes. These sessions resulted in 28 ideas reflecting three game mechanics particularly suited for educational mini-games tailored to adolescent learners.**

Keywords – mini-games; serious games; human-centred; conceptual design

I. INTRODUCTION

Nowadays, video games are no longer designed solely for entertainment purposes. The continuously increasing interest in serious games and gamification has shown that many areas can benefit from the engaging experience that video games offer. For instance, video games have been designed to help people in various therapeutic contexts [22], [56], as well as for educational purposes [47], [50], [54]. In the field of Computer-Assisted Language Learning (CALL), particularly, games have long been developed specifically for language instruction [26], [28], [45], and games have, to a more limited extent and much more recently [11], been subjected to empirical research on issues related to language development [13], [37], [42], [52].

One reason why games may be particularly suited for educational purposes is that many aspects of video games, for instance problem/puzzle solving and assessment, are also present in formal educational settings. Besides the potential of serious games to provide engaging learning experiences, previous work has shown that it is hard to bring this into practice.

More particularly, on the one hand, it has been recognized that it is challenging to ensure the effectiveness of serious games, e.g., in terms of knowledge acquisition, increased motivation or improved attitudes [41], [58]. On the other hand, many serious games have been criticized for being unable to provide compelling game experiences, thereby failing in achieving their entertaining goals [15], [29], [39].

From the perspective of the player, certainly if playing voluntary, gaming is an end in itself rather than a vehicle to learn (e.g., learn a new language) or achieve goals outside the game (e.g., live healthier) [25]. Arguably the player is ideally intrinsically motivated to play a serious game, instead of using a game to obtain an extrinsic serious goal.

Although many researchers have analysed game players' intrinsic motivations [3], [35], [36], [48], [49], [59], [61] and the design aspects that deliver engaging game experiences [21], [24], [30], [31], [32], there is still a gap in research on the design concepts and methods needed to reconcile the seemingly contradictory design goals of serious games [58].

This article addresses the challenge of how one can design serious games that are both fun and effective. In particular, the conceptual design of educational mini-games as a complementary means to the instruction in class is focused upon. Overall, mini-games are defined as small, self-contained games that usually take a short amount of time to complete and focus on a specific topic; mini-games are ubiquitous, and have been developed for several purposes, including education [17].

In this article, the typical human-centred design approach is first described in order to frame the activities of our study. Then, the research goals are specified. Further, the method and results of two conceptual design sessions are elaborated upon; this includes, firstly, the co-creation session with end-users and secondly, the brainstorm session with domain experts. Then, the results of both sessions are discussed. This article ends by summarizing the scientific contribution and delineating areas for further work.

II. RELATED WORK

The International Standardization Organization (ISO) has defined five major steps in the human-centred design process of interactive systems, in which continuous iteration is encouraged between phases of understanding and specification of the context of use (predesign), understanding and definition of user requirements (conceptual design), the production of design solutions that meet the defined requirements (design and development) and an evaluation of these designs against the requirements (evaluation*)* [27]. This process has been referred to as the "ISO9241-210 standard of Ergonomics of human-system interaction; Human-centred design for interactive systems" [27]. It relies on four basic principles, i.e., 1) active involvement of users and a good understanding of the needs of users and tasks, 2) adaptation of technology to the user, 3) iterative design, and 4) a multi-disciplinary design team.

In digital game development, human-centred design has mainly been advocated as involving players in play tests from the very moment that the first prototypes have been created [38], [43], [58]. Although this type of play testing does involve players in the evaluation phases of the (late) design process, players and stakeholders are rarely offered the opportunity to participate in predesign and (conceptual) design phases.

In excluding the eventual players from bringing in input in the early concept and design phases, game designers are likely to end up with a self-referential design, one that is oriented towards the needs and preferences of the designers rather than being tailored to the particular target group. Such an 'I'-methodology should be avoided, especially when the target audience of the game differs from the developers/designers. Besides, the game design process not only benefits from input provided by target users, but also from input given by domain experts, especially when the ambition is to reconcile entertainment and serious goals [57].

This article focuses upon a study that followed a human-centred game design process. More particularly, the results of two conceptual design sessions are reported upon. The overall research goal of the sessions was geared towards a first understanding and specification of the context of use and the definition of user requirements for educational mini-games tailored to adolescent learners. These research goals corresponded with the conceptual design phases of ISO's human-centred design process, as described earlier in the article.

As for the first conceptual design session, target users were actively involved in the idea generation of mini-game concepts for second language learning. There are several ways in which end-users can be involved in the initial phases of the design process [4], [16], [18], [34], [60]. In our study, generative techniques for idea generation were focused upon. These techniques typically rely on the creation of artefacts together with end-users, an approach that in literature is referred to as co-design or co-creation. The central notion is that the people destined to use the product play a critical role in conceptualizing and designing the product [46].

Co-creation is based upon the premise that human beings' knowledge, feelings and dreams are hard to uncover as this information may not readily be expressed in words, or cannot be observed as it might, for instance, be about latent needs. Generative techniques or 'make-tools' are needed then to facilitate the expression and communication of thoughts, feelings, and dreams. The act of physically laying out words, images and constructing representations of ideas enables the participants to articulate their ideas more thoroughly than they are able to in a typical interview or conversation. Consequently, the reflexivity through the act of creation then serves as an explanatory vehicle for their needs and ideas, not as a concrete visualization of the final design specifications [23], [44].

Although the method and the results of this first conceptual design session have been described in a previous publication, presented at the ACHI 2012 conference [1], this article provides an extended reflection by the inclusion of new empirical data that were gathered during a second conceptual design study. Conceptual design typically follows an incremental process, and hence it benefits from alternation of idea generation phases, especially when several multidisciplinary views and a variety of stakeholders are brought together.

Therefore, during the second session, domain experts - including both game designers/developers and educational experts- were asked to brainstorm and reflect upon mini-game concepts that would be transferable to a variety of domains other than language learning, including mathematics, history and geography. The development of educational games poses real challenges in terms of the return on investment. That is why the conceptualization of more generic mini-game concepts was considered to be very important because it allows mini-game concepts to be re-used for several educational domains with relatively little effort and cost.

The brainstorm session with domain experts was organized according to the evolutionary approach towards design thinking, in which phases of diverging (i.e., creating choices) and converging (making choices) are typically separated and alternated [8]. In particular, the creation of ideas in multidisciplinary teams was stimulated in order to broaden the space of possibilities.

To facilitate the idea generation process, several recombination and mutation rules and techniques were relied upon. For instance, domain experts were encouraged to express themselves visually and come up with wild ideas [53]. Additionally, the participants were encouraged to produce many ideas in order to get as many ideas from the workshop as possible. This was important because it has been found that discussing multiple brainstorming outcomes is more effective than considering a single artefact when it comes to avoiding fixation [19], richer design outcomes, better idea exploration, sharing, and group rapport [14].

III. CO-CREATION WITH END-USERS

In the following paragraphs, the first conceptual design session, the co-creation with end-users, is described in detail, including the participants and procedure, and followed by a discussion of the results.

A. Method

1) Participants

A total of fourteen adolescents participated in the co-creation session. The group was divided into two subgroups; each subgroup participated in one co-creation workshop. The first workshop was organized with eight adolescents in the morning; the other workshop, consisting of six adolescents, took place in the afternoon.

All participants were between 14 and 16 years of age, only one of the 14 participants was a girl. Twelve were in general secondary education, two participants were from technical secondary education. The eight participants of the first workshop played on average 41 minutes a day; the six participants of the second workshop played on average 1 hour 35 minutes. The participants were recruited through online forums, electronic newsletters, paper flyers and posters. Although the aim was to recruit a group of adolescents that was evenly divided in terms of gender, education and game preferences, there was an overrepresentation of boys from general secondary education who play games on a regular basis.

2) Procedure

The morning and afternoon workshop lasted each approximately three hours. Each workshop consisted of an introduction, group discussion, game design round, and a final group discussion. By following these steps, we aimed to follow the typical cognitive process of creativity closely. This process is typically divided into four or five stages, including the sensitization of the problem space, incubation, inspiration and transformation stage [5], [9], [12]. These stages will be referred to in more detail below.

Introductory round: Using a slideshow presentation, the topic at hand and the co-creation methodology were explained. Then, results from previous co-creation workshops were presented. These examples were taken from domains other than language learning, in order to prevent possible bias in the creative thinking of the participants. The introduction took around 15 minutes.

Group discussion: After the introduction, the group was split into smaller subgroups. Two researchers joined each subgroup and started a short, moderated group discussion. The aim was to better understand the envisioned context of use and the end-users, which is necessary to define requirements for the design of new products [27]. More particularly, the participants' current language learning practices -both formal and informal- and their general experience with learning through games were addressed. Additionally, this group discussion was also intended as a 'sensitizing activity', which is a typical first stage in a creative process of idea generation [5], [51]. This group discussion lasted approximately 20 minutes.

Figure 1. Co-creation session in which low-fidelity prototypes of video games were created.

Figure 2. Prototype presentation and group discussion.

Game design round: Given the time of about one hour, each subgroup was asked to come up with game concepts and create low-fidelity paper prototypes of these concepts using the available materials (see Figure 1).

The creation of at least three prototype artefacts was encouraged as it has been found that creating multiple prototypes is more effective than creating a single prototype when it comes to the design outcomes, exploration, sharing, and group rapport [14].

Note that the prototypes by no means had to be complete designs, but rather served as vehicles to express and discuss ideas, needs and preferences. Further, the participants were not constrained in the creation and conceptualization to mini-games only; they were given the opportunity to think freely about a variety of game concepts for language learning instead. Only when a group of participants had come up with two concepts that were clearly regular video games instead of mini-games, the researchers encouraged the participants to think of their next game concept as a mini-game.

When looking at the different stages of the creative process of idea generation and design thinking, the game design round resembled the third stage, inspiration [44]. In this stage, possible solutions or new insights typically occur. The incubation stage was not present in our study due to practical concerns, as the workshops were scheduled on one day. Such an incubation stage typically occurs after sensitization of the problem stage and before inspiration, and allows the participants to set the problems aside for a time.

Final group discussion: After the game design round, participants presented their prototypes to each other and the researchers (see Figure 2). Participants could ask questions, comment on the prototypes and ideas, and judge the appropriateness and potential of the presented concepts. The researchers moderated this discussion, probed for more clarification with respect to certain design choices, as well as regarding a number of pre-defined topics, questioning for instance the user-oriented and personal goals of the game concepts, the role of the teacher, or the envisioned context-of-use.

Note that the group discussion was considered as an activity that represents the final stage in the creative process, i.e., transformation [9]. This stage foresees in an evaluation of the value of ideas (e.g., group discussions) and decision (e.g., via rating) with regard to the idea selection. In our study, this transformation phase lasted about an hour.

B. Results

The co-creation workshops with end-users resulted in a total of 11 game concepts. Four of the eleven games had a multiplayer mode. Six of the games incorporated a social component, like the ability to share high scores with friends, and communicate via voice chat. The choice of the platform (computer, console, mobile) was not specified for most game concepts. Some games were thought to be more suited for a specific platform than others, with game concepts ranging from a traditional mini-game on a desktop computer, to an augmented reality game on a mobile phone. Further, the results revealed a wide range of reward mechanisms, from simple scoring systems such as traditional high scores to more complex rewarding mechanisms, whereby the player gains experience points on different levels. The participants also indicated that in-game feedback mechanisms were of considerable importance. The participants agreed that the mini-games for foreign language learning should provide some kind of feedback mechanism that helps players when they are stuck, such as for instance, a built-in translator to an in-game character that aided the player as an interpreter for foreign languages.

Overall, the results revealed two main categories of game concepts for language learning, including games for formal learning on the one hand, and games in which language serves as a means of communication on the other hand. In what follows, a more detailed overview of these two categories is given.

1) Games for formal language learning

The co-creation sessions revealed three game concepts that were aimed towards formal language learning. These games shared a focus on vocabulary, were similar in terms of immediate feedback, required limited time to play, and contained little or no narrative. The latter characteristics will be outlined based on one game concept that was developed during the co-creation workshops, namely the cannon-versus-monsters game (see Figure 3).

In this game, the player has to translate a word as quick as possible in order to prevent monsters, descending on a narrow path, to reach the player. The number of bullets a player receives depends on the length of the assigned word. For instance, a four-letter word that is correctly translated gives the player four bullets to eliminate the approaching monsters. The difficulty level of the game gradually rises with each stage, offering the player not only more challenging words to translate, but also more bullets and useful power-ups -any item that temporarily gives a character new abilities, new powers, or a statistical bonus.

Feedback in the cannons-versus-monsters game is provided immediately. Every time the player fails to translate a word, the monsters come closer to the player's home, eventually destroying it when the monsters come near enough. The player should thus try to translate as many words as possible in a correct way. When the player fails to translate a word, the consequences are instantly visible as the monsters further approach the onscreen character of the player.

The cannon-versus-monster game revolves around the relatively simple goal of keeping the monster away. By translating words correctly ammunition is earned that can be used to shoot the monsters. No further narrative or plot was provided as context for the game. The cannon-versus-monsters game concept concerned a simple and short game; it did not require a lot of time to complete. Therefore, the participants argued that the game could be played in situations in which little time is available, e.g., at home for schoolwork or even at school as part of language learning classes.

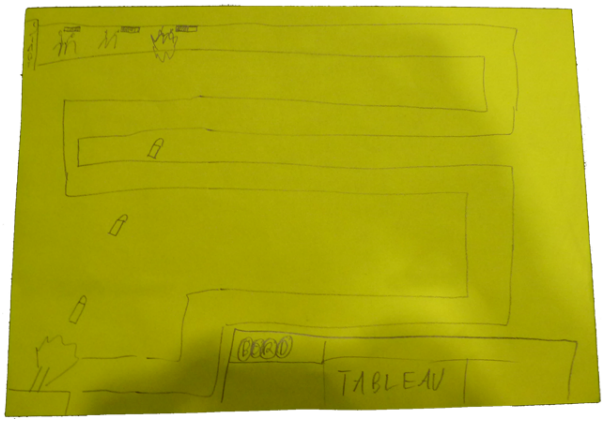


Figure 3. Drawing of the cannon-versus-monsters game concept.

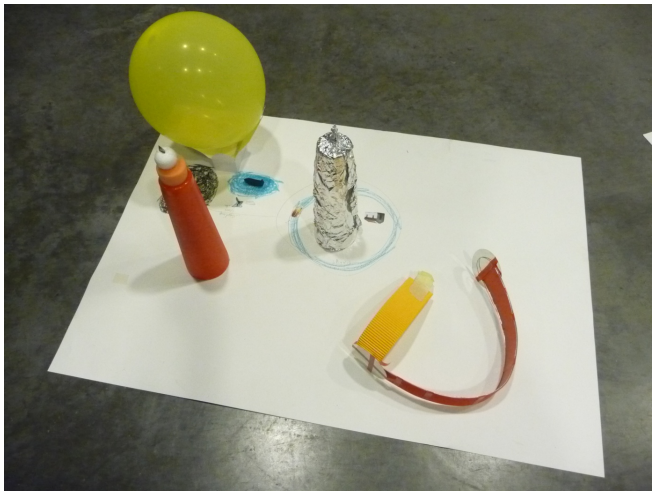


Figure 4. Creation of the adventure game.

2) Games and language as communication

The majority of ideas could be characterized as presenting a game concept in which language is used as communication. In total, six game concepts could be characterized by their focus on language as a means to communicate in the game. Players needed to communicate to progress through the game. When compared to the three game concepts that focused on vocabulary, these concepts were characterized as being more complex, containing an elaborate narrative, providing less immediate feedback, and being more time consuming. In what follows, the characteristics of the latter category of game concepts will be explained by looking at one of the six game concepts, an adventure game, in more detail (see Figure 4).

The adventure game starts from a story in which the player has to get from Paris, France, to Los Angeles, USA, to visit his or her sick mother. To achieve this, the player has to communicate with other game characters or other non-player characters. Thus, language is the means to get to the end goal. Through dialogues and creative use of language (e.g., asking for a lift, lure opponents into traps, persuasion, deceiving, etc.) the game character progresses through the game.

Overall, the results showed that the game concepts that focused on communication were more elaborate than the game concepts that focused more explicitly on linguistic sub-domains, such as vocabulary. The narrative was very important and much richer in the games focusing on communication. Consequently, feedback was thought of in a less immediate way than in the games that focused on vocabulary.

For instance, while in the cannon-versus-monsters game, the player immediately receives bullets to keep the monsters away, or sees the monsters approaching further after each mistake; the progress in the games focusing on communication was less immediately visible. Although the end goal was clear, the player was only considered to slowly approach the goal; and in this process the rewards were more high level.

Finally, compared to the games focusing on vocabulary, the game concepts with a focus on communication were relatively complex and therefore required more time to play. To engage in the game concepts, players would need a period of uninterrupted time available to play. This would make these games, according to our participants, more suited for playing at home, and less suited for class use.

IV. BRAINSTORM WITH DOMAIN EXPERTS

In the next paragraphs, the participants and procedure of the brainstorm session with domain experts is described, followed by a discussion of the results.

A. Method

1) Participants

Three game designers/developers and three educational experts were invited to participate in the brainstorm session. These domain experts were divided in three subgroups of two people, one game expert and one educational expert. Figure 5 shows an impression of the brainstorm session in small subgroups.

2) Materials

The domain experts were provided with three kinds of materials as input for the workshop in order to facilitate the brainstorming process; they were not forced to use these materials.

Firstly, each subgroup received five random cards from the Game Seeds© card deck [55] (see Figure 6). These cards visually represent game mechanics as well as some playful rules that can be used to turn a brainstorming exercise into a playful activity. Even though the participants did not have to engage in the entire Game Seeds© brainstorming game, the visual presentation of game mechanics on the cards was considered to impose additional constraints that could increase creative thinking.

Secondly, to encourage domain experts to generate game concepts for a wide range of educational purposes, sheets with written topics that typically constitute the subject matter of adolescent-oriented class courses on geography, history, mathematics and language learning were handed over. Again, these sheets were shown for inspiration purposes only; domain experts did not need to rely on this information.

Thirdly, personas were given as input to the brainstorm in order to further improve the idea generation process. Personas represent fictive characters that are based on factual information. In human-centered design, personas are utilized to present a reflection of the (hypothetical) archetypal user. Additionally, personas are also considered useful as a method to make a design team empathise with the product's end-users thanks to a deeper understanding of their likes and dislikes, and their capabilities [2], [10], [40].

Figure 5. Brainstorming session with domain experts.

Figure 6. Game Seeds© cards, used as input for the brainstorming method. The visual representation of concrete game mechanics presented constraints that were deliberately imposed to increase the likelihood of creative design thinking.

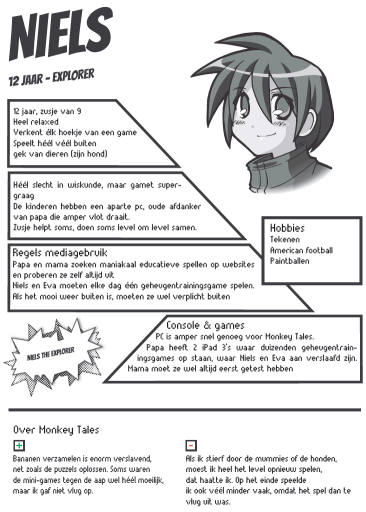


Figure 7. Extreme persona (Dutch), used as input for the brainstorm session with domain experts, as a means to create empathy with the end users and stimulate creativity during the idea generation process.

In our study, the personas resulted from the insights gained in a diary study that preceded the brainstorming session. In the diary study, eight households with adolescents were selected to report on their behaviour and experiences with Monkey Tales©, an educational game for training mathematics [33], for a period of two to three weeks. The results were summarized in 'extreme' personas that enlarged the players' general characteristics and reported in an exaggerated way their likes and dislikes, the media rules of the household, their game preferences and experiences. For an example of such an extreme personas (in Dutch), we refer to Figure 7.

The advantages of brainstorming with extreme inputs that are based on evidence-based personas are twofold. First, personas have been found to provide a powerful means to communicate relevant user characteristics to help a product design team to empathize more with the envisioned end-users [2], [10], [40]. Secondly, the power of the reflection on extreme stimuli has been recognized as showing great promise to increase the creativity during idea generation phases in user-centred design [6], [7], [19].

3) Procedure

The procedure consisted of three rounds of brainstorming in pairs followed by evaluative, plenary discussions of the value of the ideas. After each brainstorming and discussion round, new pairs were formed and a new round of brainstorming and evaluative discussion was started. In each round, the dyads received a new combination of input materials, including a persona, five randomly assigned game mechanics and a new educational topic.

In total, the brainstorm lasted approximately two hours and 30 minutes.

B. Results

The brainstorm session with domain experts generated 28 ideas that reflected three different types of educational mini-game concepts. The first category concerned the 'Matchers', characterized as those mini-game concepts in which players are challenged to combine several related things such as words, numbers, images or topics. The majority of the game concepts generated during the brainstorm session could be classified as a Matcher. The second category encompassed the "Sorters". These concepts shared the common idea that things can be ordered on a timeline or map. The third main category of game concepts revolved around Multiple Choice (questions). It should be noted, though, that some game concepts could not be classified into one of these three aforementioned game categories. In sum, the analysis of the game concepts revealed the following classification: a) Matchers, b) Sorters, c) Multiple Choice and d) Others.

The constraints of this article do not allow us to elaborate upon each single idea that was generated during the brainstorm. Consequently, we will select one game concept for each category and discuss it in more detail.

1) Matchers

In total, the brainstorm session resulted in 16 Matcher game concepts. To exemplify this, we will report upon the 'Snowlines' concept. Figure 8 shows the concept drawing as it was sketched and presented during the brainstorm session.

The idea of 'Snowlines' is that by snowboarding/skating/skiing through the gates the player needs to select the words or items that are related to each other. There are several routes that are allowed/correct; some are more optimal than others and by choosing or combining the right gates, more or less points can be collected. The learning concept relies on the ability to group similar words, items, topics, verb conjunctions etc. The underlying game mechanic is based on selection and grouping. For this concept, one might think of additional features to enhance the playability, such as doing tricks in between two gates.

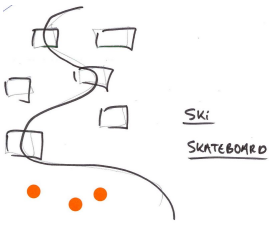


Figure 8. Concept sketch 'Snowlines', a Matcher game concept.

2) Sorter

Four Sorter game concepts were generated. One of these Sorters was the 'Character-Map Exploration' game, depicted in Figure 9. In this game concept, players would receive an inventory of characters, of which some are locked, and some are not. The players would then need to interview these characters to get to know them. Based on the information revealed by the characters, players would be able to place the characters on the correct spot on the map. According to the domain experts, linking characters and their items with the corresponding country and locating this country, the players might eventually learn about historical people, their stories and locations.

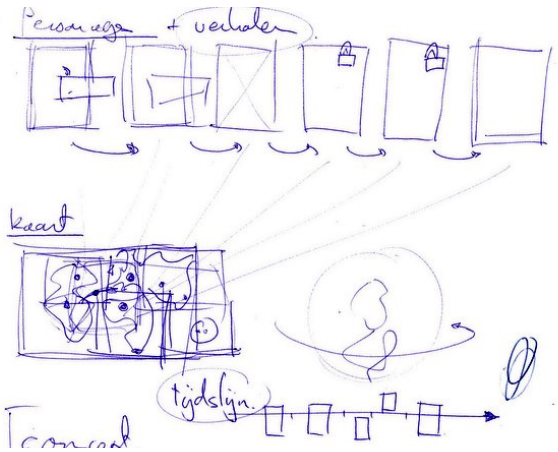


Figure 9. Concept sketch 'Character-Map Exploration', a Sorter game.

3) Multiple Choice

The brainstorm session resulted in two multiple Choice game concepts. Figure 10 shows the concept sketch of one of these, namely the 'Save the Princess' game. The game concept resolved around the challenge of saving a princess and finding your way through a labyrinth. By choosing the right door/way, the player would select the correct answer. Behind the wrong doors/answers, there is the risk to be confronted with deadly monsters. Everywhere in the labyrinth the player might engage in finding clues and hints to help find the correct way.

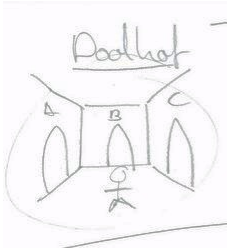


Figure 10. Concept sketch 'Save the Princess', a Multiple Choice game.

4) Others

Six concepts could not be classified in one of the categories above. Examples were the 'Search the 7 Mistakes' in which players had to select things that are wrong in an image, and the 'Draw and Guess' game concept in which players would have to guess the specific word from his or her teammates' drawings. These two game concepts are represented in Figure 11.

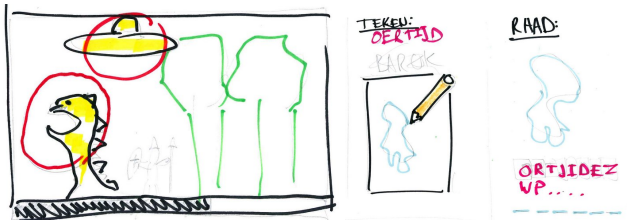


Figure 11. Concept sketches 'Search the 7 mistakes' (left) and 'Draw and Guess' (right).

V. DISCUSSION

A variety of ideas were generated during the brainstorming session with domain experts and during the co-creation workshops with end-users, showing considerable promise on the basis of which both fun and effective educational mini-games can be designed.

The results of the co-creation with end-users point towards two directions that can be chosen in the design phase of educational games for language learning. On the one hand, the participants considered the potential of games for formal language learning that resembled the definition of a mini-game. It should hereby be noted that these formal language-learning exercises were related to learning vocabulary; none of these game concepts dealt with grammar. It is unclear, however, whether this is the result of a lack of interest or the adolescent participants' lack of capability to consider grammar-related game concepts. On the other hand, when it comes to language as communication, the participants preferred more complex games with a narrative, which confirms that games can be used as a medium to create a need for the language learner to accomplish objectives that lie outside the language itself. Nevertheless, many of this type of game concepts resembled existing games. It is not clear whether this fixation upon existing game concepts resulted from methodological decisions, the participants' characteristics, or from the combination of both aspects. In order to potentially be able to answer this question, we would advise future research to link the participants' gaming history and pre-existing preferences to the creation of game concepts in co-creation sessions and experiment with a wider variety of idea generation techniques.

Secondly, when it comes to the ideation of mini-game concepts that can be used for several exercises in a variety of instructional domains other than language learning only, the results suggested that Matchers, Sorters and Multiple Choice game mechanics are most promising to be included in educational mini-games. We are convinced that the generic character of these game concepts may bring an important advantage for game developers or publishers who want to wisely invest in educational mini-games.

There are some issues that remain unsolved, though. For instance, it is not clear why the brainstorm session mainly revealed Matcher concepts. At this point, it is unclear whether this is due to a lack of a better classification, or whether it is just easier to generate ideas about Matcher game concepts; or if there are even other reasons involved. Moreover, it should be noted, that the categorization of Matchers, Sorters and Multiple Choice educational game mechanics does not differ that much from the classical e-learning approaches that also typically rely on matching, multiple choice and sorting. In this context, our results clearly indicated that the 'packaging' of the exercises provided the mini-games with an additional layer of fantasy that increased the game experience (e.g., story, characters and missions.

Another issue with respect to the brainstorm session with domain experts concerns the usefulness of the input materials, i.e., the personas, the Game Seeds© and the instructional examples. To our knowledge, there is no previous work in which the combination of these input materials have been employed. Consequently, it is unclear what the effect is of the methodological decisions upon the brainstorm outcome.

The last issue of our discussion is relevant for both the co-creation and brainstorming session. It should be stressed once more that the results were not intended to represent finished game concepts or concrete design guidelines. The results provide the design team with more insights into the users and their preferences, information that should be complemented with for instance the insights revealed through contextual observations. When the design team understands the users and the context of use, inspiration can be drawn from the created artefacts (co-creation workshops) and presented sketches (brainstorm session) in the further development from low-fidelity to high-fidelity prototypes. As the ISO 9241-210 human-centred design process prescribes, phases of development and human-based evaluations should sufficiently iterate in order to optimize the end product. Hence, the designers can put their own expertise in developing the designs, being inspired but at the same time not limited by the artefacts that were created during the brainstorm and co-creation sessions. The design team should hereby acknowledge that the products have to be re-evaluated in several iterations by directly involving the stakeholders, i.e., the end-users in the first place.

VI. CONCLUSION AND FUTURE WORK

This article reported upon two conceptual design sessions in which a human-centred approach was followed to gather requirements and inspiration for the design of mini-games with educational purposes tailored to adolescent learners.

Firstly, co-creation workshops were held with adolescents in order to reveal their ideas, needs and preferences with regard to video games for language learning. The results showed a divide between the concepts for mini-games that were oriented towards formal language learning (e.g., exercises on vocabulary) on the one hand and video games that were based on communication with others (players or in-game characters) on the other.

Secondly, brainstorm sessions were held with domain experts, including game designers/developers and educational experts, to generate ideas and gather requirements for the design of mini-game concepts with educational goals. The results revealed a categorization of educational mini-game concepts with sufficient potential to be both fun and efficient, including Matchers, Sorters and Multiple-Choice concepts. The Matchers seemed most promising, not only did this category generate most ideas, it also turned out to be most promising to be applied in a variety of educational programs ranging from mathematics to language learning, geography and history.

To conclude, the two conceptual design sessions described in this article resulted in a divergence and multitude of rich ideas. Nevertheless, more design iterations are needed to evaluate these ideas by making choices in terms of the most promising ideas. By no means are the conceptual design sessions imposing final solutions. As described by the ISO's human-centred design process, more iterations and empirical evaluations are needed in the subsequent detailed design and development phases. Consequently, future work should focus on the next human-centred design steps and report which design decisions are to be taken to realize successful educational mini-games that are tailored to adolescent learners; mini-games that reconcile both entertaining and educational goals.

ACKNOWLEDGMENT

The first study (co-creation sessions) was conducted as part of the iMinds-MiGaMe (Mini Games for Maximum efficiency) project. The second study (the brainstorm sessions with domain experts) was performed in the context of the iMinds-G@S (Games at School) project. These projects are cofounded by iMinds (Interdisciplinary Institute for Technology), a research institute founded by the Flemish Government. The leading companies involved in the iMinds-G@S project are Larian and die Keure, with project support of IWT. We are grateful to all participants of the co-creation and brainstorming sessions for their enthusiasm to generate cool ideas.

REFERENCES

[1] Poels, Y., Annema, J.-H., Zaman, B., and Cornillie, F. "Developing User-Centered Video Game Concepts for Language Learning," ACHI conference, 2012, pp. 11–16.

[2] Antle, A.N. Child-personas: fact or fiction?" Proceedings of the 6th conference on Designing Interactive systems (DIS 06), ACM, 2006, pp. 22–30.

[3] Bartle, R. "Hearts, Clubs, Diamonds, Spades: Players Who Suit MUDS," 1996, http://www.mud.co.uk/richard/hcds.htm [last accessed Dec 2012].

[4] Beyer, H. and Holtzblatt, K. "Contextual Design: Defining Customer-centered Systems," Morgan Kaufmann, San Francisco, CA, 1998.

[5] Boden, M.A. "Creative Mind: Myths and Mechanisms," Routledge, New York, NY, 2003.

[6] Bowen, S. "Getting it Right: Lessons Learned in Applying a Critical Artefact Approach," In: Undisciplined! Design Research Society Conference 2008, Sheffield Hallam University, Sheffield, UK, 2009.

[7] Bowen, S.J. "Crazy ideas or creative probes?: presenting critical artefacts to stakeholders to develop innovative product ideas," Proceedings of EAD07: Dancing with Disorder: Design, Discourse and Disaster, Izmir, Turkey, 2007.

[8] Brown, T. "Change by Design: How Design Thinking Transforms Organizations and Inspires Innovation," HarperBusiness, 2009.

[9] Bullinger, H.-J., Müller-Spahn, F., and Rössler, A. "Encouraging Creativity - Support of Mental Processes by Virtual Experience," Conference Virtual Reality World, 1996.

[10] Cooper, A. and Reimann, R.M. "About Face 2.0: The Essentials of Interaction Design," Wiley, 2003.

[11] Cornillie, F., Thorne, S.L., and Desmet, P. "Editorial. Digital games for language learning: from hype to insight?" ReCALL vol. *24*(3), 2012, pp. 243–256.

[12] Csikszentmihalyi, M. "Creativity: Flow and the Psychology of Discovery and Invention," Harper Perennial, 1997.

[13] deHaan, J., Reed, W.M., and Kuwada, K. "The effect of interactivity with a music video game on second language vocabulary recall," Language Learning & Technology vol. 14(2), 2010, pp. 74–94.

[14] Dow, S., Fortuna, J., Schwartz, D., Altringer, B., Schwartz, D., and Klemmer, S. "Prototyping dynamics: sharing multiple designs improves exploration, group rapport, and results," Proceedings of the 2011 annual conference on Human factors in computing systems, (CHI2011), ACM, 2011, pp. 2807–2816.

[15] Egenfeldt-Nielsen, S. "Beyond Edutainment: Exploring the Educational Potential of Computer Games." PhD Dissertation, online http://www.learninginvideogames.com/research-and-papers/beyond-edutainment-a-dissertation-by-simon-egenfeldt-nielsen/ [last accessed December 2012], 2005.

[16] Ermi, L. and Mäyrä, F. "Player-Centred Game Design: Experiences in Using Scenario Study to Inform Mobile Game Design," Game Studies: The International Journal of Computer Game Research vol. 5(1), 2005.

[17] Frazer, A., Argles, D., and Wills, G. "Assessing The Usefulness Of Mini-games As Educational Resources," ALT-C 2007: Beyond Control, 2007.

[18] Gaver, B., Dunne, T., and Pacenti, E. "Design: Cultural probes," Interactions vol. 6(1), 1999, pp. 21–29.

[19] Gaver, B. and Martin, H. "Alternatives: exploring information appliances through conceptual design proposals," Proceedings of the SIGCHI conference on Human factors in computing systems (CHI2000), 2000, pp. 209–216.

[20] Gaver, W.W., Beaver, J., and Benford, S. "Ambiguity as a resource for design," Proceedings of the SIGCHI conference on Human factors in computing systems (CHI2003), ACM, 2003, pp. 233–240.

[21] Greenberg, B.S., Sherry, J., Lachlan, K., Lucas, K., and Holmstrom, A. "Orientations to Video Games Among Gender and Age Groups, "Simulation & Gaming vol. 41(2), 2010, pp. 238–259.

[22] Griffiths, M. "The therapeutic use of videogames in childhood and adolescence, "Clinical child psychology and psychiatry vol. 8(4), 2003, p.547.

[23] Hackos, J.A. and Redish, J.C. "User and Task Analysis for Interface Design, "John Wiley & Sons, Inc., 1998.

[24] Hartmann, T. and Klimmt, C. "Gender and Computer Games: Exploring Females' Dislikes," Journal of Computer-Mediated Communication vol. 11(4), 2006, pp. 910–931.

[25] Hubbard, P. "Evaluating Computer Games for Language Learning," Simulation & Gaming vol. 22(2), 1991, pp. 220 – 223.

[26] Hubbard, P. "Interactive Participatory Dramas for Language Learning," Simulation & Gaming vol. 33(2), 2002, pp. 210 – 216.

[27] ISO. "ISO 9241-210:2010 - Ergonomics of human-system interaction -- Part 210: Human-centred design for interactive systems," http://www.iso.org/iso/catalogue_detail.htm?csnumber=52075 [last accessed September 2012], 2010.

[28] Johnson, W.L. "Serious Use of a Serious Game for Language Learning," Proceeding of the 2007 conference on Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work, IOS Press 2007, pp. 67–74.

[29] Kirriemuir, J. and McFarlane, A. "Literature review in games and learning," Futurelab, 2004.

[30] Klimmt, C., Hartmann, T., and Frey, A. "Effectance and control as determinants of video game enjoyment," Cyberpsychology & Behavior vol. 10(6), 2007, pp. 845–847.

[31] Klimmt, C., Rizzo, A., Vorderer, P., Koch, J., and Fischer, T. "Experimental Evidence for Suspense as Determinant of Video Game Enjoyment." Cyberpsychology & Behavior vol. 12(1), 2009, pp. 29–31.

[32] Klimmt, C., Schmid, H., and Orthmann, J. "Exploring the Enjoyment of Playing Browser Games." Cyberpsychology & Behavior vol. 12(2), 2009, pp. 231–234.

[33] Larian Studios and Die Keure. "Monkey Tales Games - Educational math games." http://www.monkeytalesgames.com [last accessed: September 2012], 2011.

[34] Laurel, B. and Lunenfeld, P. "Design Research: Methods and Perspectives," The MIT Press, 2003 .

[35] Lazzaro, N. "Why We Play Games: Four Keys to More Emotion Without Story," Game Developers Conference, 2004.

[36] Malone, T.W. and Lepper, M.R. "Making learning fun: A taxonomy of intrinsic motivations for learning," Aptitude, learning, and instruction vol. 3, 1987, pp. 223–253.

[37] Miller, M. and Hegelheimer, V. "The SIMs meet ESL. Incorporating authentic computer simulation games into the language classroom " Interactive Technology and Smart Education vol. 3(4), 2006, pp. 311–328.

[38] Pagulayan, R.J., Keeker, K., Wixon, D., Romero, R.L., and Fuller, T. "User-centered design in games," In J.A. Jacko and A. Sears, eds. The Human-Computer Interaction Handbook. Fundamentals, Evolving Technologies and Emerging Applications. L. Erlbaum Associates Inc., Hillsdale, NJ, USA, 2003, pp. 883–906.

[39] Papert, S. "Does Easy Do It? Children, Games, and Learning *"* Game Developer, 1998, pp. 87–88.

[40] Pruitt, J. and Grudin, J. "Personas: practice and theory". Proceedings of the 2003 conference on Designing for user experiences (DUX '03), ACM 2003, pp. 1–15.

[41] Purushotma, R., Thorne, S.L., and Wheatley, J. "10 Key Principles for Designing Video Games for Foreign Language Learning," 2008. http://knol.google.com/k/10-key-principles-for-designing-video-games-for-foreign-language-learning# [last accessed May 2011].

[42] Ranalli, J. "Learning English with The Sims: exploiting authentic computer simulation games for L2 learning," Computer Assisted Language Learnin, vol.21(5), 2008, pp. 441–455.

[43] Salen, K. and Zimmerman, E. "Rules of Play: Game Design Fundamentals," MIT Press, 2003.

[44] Sanders, E.B.N. "From user-centered to participatory design approaches," Design and the social sciences: making connections vol. 2, 2002, p. 1.

[45] Sanders, R.H. and Sanders, A.F. "History of an AI spy game: Spion," In Thirty years of computer assisted language instruction. CALICO, 1995, pp. 114–127.

[46] Schuler, D. and Namioka, A. "Ethnographic Field Methods and Their Relation to Design," In Participatory Design: Principles and Practices. Routledge, 1993.

[47] Serious Games Interactive. "Global Conflicts: Palestine," Gamers Gate, Manifesto Games & Macgamestore, 2007.

[48] Sherry, J.L., Lucas, K., Greenberg, B.S., and Lachlan, K. "Video game uses and gratifications as predictors of use and game preference," In Playing video games: Motives, responses, and consequences, Lawrence Erlbaum Associates, 2006, pp. 213–224.

[49] Sherry, P., Lucas, K., Greenberg, B., and Lachlan, K. "Playing video games: motives, responses, and consequences, "Routledge, 2006.

[50] Spongelab Interactive. "Genomics Digital Lab - History of biology," Spongelab Interactive, 2009.

[51] Sulmon, N., Derboven, J., Montero Perez, M., and Zaman, B. "Creativity as a process: a participatory design approach to gather mobile language learning user requirements," Proceedings of the IADIS international conference e-learning 2011, 2011, pp. 429–437.

[52] Sykes, J.M. "Learner requests in Spanish: Examining the potential of multiuser virtual environments for L2 pragmatic acquisition," In L. Lomicka and G. Lord, eds. The Next Generation: Social Networking and Online Collaboration in Foreign Language Learning. Computer Assisted Language Instruction Consortium, Durham, 2009.

[53] Thoring, K. and Müller, R.M. "Understanding the creative mechanisms of design thinking: an evolutionary approach, " Proceedings of the Second Conference on Creativity and Innovation in Design (DESIRE'11), ACM, 2011, pp. 137–147.

[54] United Nations World Food Programme. "Food Force: The First Humanitarian Video Game," World Food Programme, 2005.

[55] Utrecht School of the Arts, Monobanda, and Metagama. "GAME SEEDS" 2012. http://www.gameseeds.net/ [last accessed September 2012].

[56] Vanden Abeele, V., Geurts, L., Husson, J., Windey, F., Annema, J., Verstrate, M., and Desmet, S. "Designing Slow Fun! Physical Therapy Games to Remedy the Negative Consequences of Spasticity," Proceedings of the 3rd International Conference on Fun and Games (FnG 2010), ACM Press, 2010.

[57] Vanden Abeele, V. and De Schutter, B. "Designing intergenerational play via Enactive Interaction," Competition and Acceleration. Journal of Personal and Ubiquitous Computing: Special Issue on Design for Social Interaction through Physical Pla*y*, 2009.

[58] Vanden Abeele, V., Schutter, B., Geurts, L., Desmet, S., Wauters, J., Husson, J., Van Audenaeren, L., Broeckhoven, F., Annema, J.,and Geerts, D. "P-III: A Player-Centered, Iterative, Interdisciplinary and Integrated Framework for Serious Game Design and Development," In S. Wannemacker, S. Vandercruysse and G. Clarebout, eds., Serious Games: The Challenge. Springer Berlin Heidelberg, 2012, pp. 82–86.

[59] Vorderer, P. and Bryant, J. "Playing Video Games: Motives, Responses, and Consequences," Lawrence Erlbaum Associates, 2006.

[60] Winograd, T. and Kuhn, S. "Participatory Design,"In Bringing Design to Software. Addison-Wesley, 1996.

[61] Yee, N. "Motivations for Play in Online Games." CyberPsychology & Behavior vol. 9(6), 2006, pp. 772–775.

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, ENERGY, COLLA, IMMM, INTELLI, SMART, DATA ANALYTICS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING, MOBILITY, WEB

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO, SOTICS, GLOBAL HEALTH

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION, VEHICULAR, INNOV

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS, CLOUD COMPUTING, COMPUTATION TOOLS, IMMM, MOBILITY, VEHICULAR, DATA ANALYTICS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL, INFOCOMP

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA, COCORA, PESARO, INNOV

issn: 1942-2601