



ICIW 2011

The Sixth International Conference on Internet and Web Applications and Services

ECCSS 2011

The First International Workshop on Enterprise Cloud Computing
- Strategies and Solutions -

March 20-25, 2011

St. Maarten, The Netherlands Antilles

ICIW 2011 Editors

Mihhail Matskin, NTNU, Norway

Mark Perry, University of Western Ontario - London, Canada

Zaigham Mahmood, University of Derby, UK

ICIW 2011

Foreword

The Sixth International Conference on Internet and Web Applications and Services (ICIW 2011) held on March 20-25, 2011 in St. Maarten, The Netherlands Antilles, continued a series of co-located events that covered the complementary aspects related to designing and deploying of applications based on IP&Web techniques and mechanisms.

Internet and Web-based technologies led to new frameworks, languages, mechanisms and protocols for Web applications design and development. Interaction between web-based applications and classical applications requires special interfaces and exposes various performance parameters.

Web Services and applications are supported by a myriad of platforms, technologies, and mechanisms for syntax (mostly XML-based) and semantics (Ontology, Semantic Web). Special Web Services based applications such as e-Commerce, e-Business, P2P, multimedia, and GRID enterprise-related, allow design flexibility and easy to develop new services. The challenges consist of service discovery, announcing, monitoring and management; on the other hand, trust, security, performance and scalability are desirable metrics under exploration when designing such applications.

ICIW 2011 comprised five complementary tracks. They focused on Web technologies, design and development of Web-based applications, and interactions of these applications with other types of systems. Management aspects related to these applications and challenges on specialized domains were aided at too. Evaluation techniques and standard position on different aspects were part of the expected agenda.

ICIW 2011 also included:

- ECCSS 2011, The First International Workshop on Enterprise Cloud Computing - Strategies and Solutions

We take this opportunity to thank all the members of the ICIW 2011 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the ICIW 2011. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICIW 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICIW 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in Web Services.

We are convinced that the participants found the event useful and communications very open. The beautiful places of St. Maarten surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

ICIW 2011 Chairs

Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany

Mihhail Matskin, NTNU, Norway

Guadalupe Ortiz, University of Cádiz, Spain

Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada

Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan / New York State Bar, NY, USA

Javier Cubo, University of Malaga, Spain

Didier Sebastien, University of Reunion Island, France

Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada

Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan / New York State Bar, NY, USA

Guadalupe Ortiz, University of Cádiz, Spain

Natalia Kryvinska, University of Vienna, Austria

Jean-Pierre Gerval, ISEN Brest, France

Matthias Ehmann, University of Bayreuth, Germany

Benoit Christophe, Alcatel-Lucent Bell Labs, France

ECCSS 2011 Chairs

Zaigham Mahmood, University of Derby, UK

Zhengxu Zhao, Shijiazhuang Tiedao University, China

ICIW 2011

Committee

ICIW Advisory Chairs

Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Mihhail Matskin, NTNU, Norway
Guadalupe Ortiz, University of Cádiz, Spain
Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada
Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan / New York State Bar, NY, USA

ICIW Publicity Chairs

Javier Cubo, University of Malaga, Spain
Didier Sebastien, University of Reunion Island, France

Special Areas Chairs

SLAECE Chairs

Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada

SERCOMP Chairs

Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan / New York State Bar, NY, USA
Guadalupe Ortiz, University of Cádiz, Spain
Natalia Kryvinska, University of Vienna, Vienna

VEWAeL Chair

Jean-Pierre Gerval, ISEN Brest, France
Matthias Ehmann, University of Bayreuth, Germany

SERCOMP Chairs

Benoit Christophe, Alcatel-Lucent Bell Labs, France

ICIW 2011 Technical Program Committee

Silvana Vanesa Aciar, Universidad Nacional de San Juan (UNSJ), Argentina
Mehmet S. Aktas, TUBITAK (Turkish National Science Foundation), Turkey
Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Markus Alekxy, ABB Corporate Research Center, Germany
Jameela Al-Jaroodi, College of Information Technology, UAEU
Giner Alor Hernandez, Instituto Tecnológico de Orizaba - Veracruz, México
Feda Alshahwan, The University of Surrey, UK
Eckhard Ammann, Reutlingen University, Germany
José Antonio Mateo, University of Castilla-La Mancha - Albacete, Spain
Marzieh Asgarnezhad, Islamic Azad University of Kashan, Iran
Siegfried Benkner, University of Vienna, Austria

Giancarlo Bo, Technology and Innovation Consultant- Genova, Italy
Luis Borges Gouveia, University Fernando Pessoa - Porto, Portugal
Christos Bouras, University of Patras / Research Academic Computer Technology Institute, Greece
Laure Bourgois, INRETS, France
Mahmoud Brahim, University of Msila, Algeria
Dung Cao, University of Bordeaux 1 - Talence, France
Miriam A. M. Capretz, The University of Western Ontario - London, Canada
Serge Caumette, University Bordeaux 1, France
Ajay Chakravarthy, University of Southampton, UK
Xi Chen, Nanjing University, China
Benoit Christophe, Alcatel-Lucent Bell Labs, France
Sam Chung, University of Washington - Tacoma, USA
Marco Comuzzi, Eindhoven University of Technology, The Netherlands
Javier Cubo, University of Malaga, Spain
Alfredo Cuzzocrea, Italian National Research Council / University of Calabria, Italy
Paulo da Fonseca Pinto, Universidade Nova de Lisboa, Portugal
María del Pilar Villamil, Universidad de los Andes - Bogotá, Colombia
Gregorio Diaz Descalzo, University of Castilla - La Mancha, Spain
Eugeni Dodonov, Mandriva, Brazil
Ioan Dzitac, Aurel Vlaicu University of Arad, Romania
Julian Eckert, TU-Darmstadt, Germany
Javier Fabra, University of Zaragoza, Spain
Jacques Fayolle, Télécom Saint-Etienne/l'Université Jean Monnet, France
Mário Freire, University of Beira Interior, Portugal
Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Ingo Friese, Deutsche Telekom AG - Berlin, Germany
Xiang Fu, Hofstra University, USA
Roberto Furnari, Università di Torino, Italy
Stefania Galizia, Innova, Italy
Ivan Ganchev, University of Limerick, Ireland
G.R. Gangadharan, NOVAY, The Netherlands
Bozhidar Georgiev, Technical University of Sofia, Bulgaria
Olivier Gerbe, HEC, Canada
Jean-Pierre Gerval, ISEN Brest, France
Mohamed Gharzouli, Mentouri University of Constantine, Algeria
Gustavo González-Sánchez, Mediapro Research, Spain
Feliz Gouveia, Universidade Fernando Pessoa - Porto, Portugal
Anna Goy, Università di Torino, Italy
Bidyt Gupta, Southern Illinois University - Carbondale, USA
Ileana Hamburg, Institut Arbeit und Technik, Germany
Rattikorn Hewett, Texas Tech University, USA
Dag Hovland, Universitetet i Bergen, Norway
Anna Hristoskova, Ghent University, Belgium
Chi Chi Hung, Tsinghua University - Beijing, China
Edward Hung, Honk Kong Polytechnic University, Hong Kong
Omar Hussain, Curtin University of Technology, Australia
Giovambattista Ianni, Università della Calabria, Italy
Linda Jackson, Michigan State University, USA

Nicolas James, Ecole Centrale Paris, France
Ivan Jelinek, Czech Technical University, Czech Republic
Carlos Juiz, Universitat de les Illes Balears - Palma de Mallorca, Spain
Monika Kaczmarek, Poznan University of Economics, Poland
Herman Kaindl, TU-Wien, Austria
Jalal Karam, Alfaisal University - Riyadh, Kingdom of Saudi Arabia
Rachid El Abdouni Khayari, The University of the German Federal Armed Forces - Munich, Germany
Brigitte Kerherve, UQAM, Canada
Suhyun Kim, Korea Institute of Science and Technology (KIST), Korea
K.P. Lam, University of Keele, UK
Federica Landolfi, University of Sannio, Italy
HoonKi Lee, Electronics & Telecommunications Research Institute(ETRI), Republic of Korea
Ho-fung Leung, The Chinese University of Hong Kong, China
Longzhuang Li, Texas A&M University-Corpus Christi, USA
Shiguo Lian, France Telecom/Orange Labs -Beijing, China
Lu Liu, Middlesex University, UK
Malamati Louta, University of Western Macedonia - Kozani, Greece
Zoubir Mammeri, IRIT - Toulouse, France
Chengying Mao, Jiangxi University of Finance and Economics, China
Bertrand Mathieu, Orange Labs - Lannion France
Mihhail Matskin, NTNU, Norway
Hamid Mcheick, Université du Québec à Chicoutimi, Canada
Imen Tayari Meftah, Polytech Nice Sophia Antipolis, France
Panagiotis Takis Metaxas, Wellesley College, USA
Ralph Mietzner, University of Stuttgart, Germany
Fernando Miguel Carvalho, Lisbon Superior Engineering Institute, Portugal
Nader Mohamed, College of Information Technology/UAE University, United Arab Emirates
Shahab Mokarizadeh, Royal Institute of Technology (KTH), Sweden
Christof Momm, SAP AG Research Center CEC Karlsruhe, Germany
Francisco Montero, University of Castilla - La Mancha, Spain
Gero Mühl, University of Rostock,, Germany
Kaninda Musumbu, University Bordeaux 1 - Talence, France
Alex Ng, The University of Ballarat, Australia
Ulrich Norbistrath, University of Tartu, Estonia
Theodoros Ntouskas, Univeristy of Piraeus, Greece
Jason R.C. Nurse University of Warwick - Coventry, UK
Asem Omari, University of Hail, Kingdom of Saudi Arabia
Guadalupe Ortiz, University of Cádiz , Spain
Carol Ou, The Hong Kong Polytechnic University - Kowloon, Hong Kong
Helen Paik, University of New South Wales - Sydney, Australia
Apostolos Papageorgiou, Technische Universitaet Darmstadt, Germany
João Paulo Sousa, Instituto Politécnico de Bragança, Portugal
Fredrik Paulsson, Umeå University, Sweden
George Pentafronimos, University of Pireaus, Greece
Stefan Pietschmann, Technische Universität Dresden, Germany
Dorin-Mircea Popovici, Universitatea "Ovidius" Constanța, Romania
Marc Pous Marín, Barcelona Digital Centre Tecnològic, Spain
Idris A. Rai, Makerere University, Uganda

Lucia Rapanotti, The Open University - Milton Keynes, UK
José Raúl Romero, Universidad de Cordoba/Campus de Rabanales, Spain
Christoph Reinke, University of Lübeck, Germany
Nicolas Repp, Technische Universität Darmstadt, Germany
Daniele Riboni, Università degli Studi di Milano, Italy
Jan Richling, Technical University Berlin, Germany
Sebastian Rieger, Karlsruhe Institute of Technology (KIT) / Steinbuch Centre for Computing (SCC), Germany
Miguel Rojas, TU-Dortmund, Germany
Giancarlo Ruffo, Università degli Studi di Torino, Italy
Antonio Ruiz Martínez, University of Murcia, Spain
Fatiha Sadat, Université du Québec à Montréal, Canada
Sébastien Salva, IUT d'Aubière, France
Brahmananda Sapkota, University of Twente, The Netherlands
Monika Schubert, Graz University of Technology, Austria
Stefan Schulte, Technische Universität Darmstadt, Germany
Didier Sebastien, University of Reunion Island, France
Véronique Sébastien, University of Reunion Island, France
Jawwad Shamsi, National University of Computer & Emerging Sciences - Karachi, Pakistan
Lijie Sheng, Xidian University - Xi'an, China
Patrick Siarry, Université Paris 12 (LiSSi) - Créteil, France
Florian Skopik, Vienna University of Technology, Austria
Vladimir Stanchev, Asperado Ltd., Germany
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Vladimir Šor, University of Tartu, Estonia
Sayed Gholam Hassan Tabatabaei, Universiti Teknologi Malaysia (UTM) - Kuala Lumpur, Malaysia
Nazif Cihan Tas, Siemens Corporate Research - Princeton, USA
Vagan Terziyan, University of Jyväskylä, Finland
Pierre Tiako, Langston University - Oklahoma, USA
Leonardo Tininini, ISTAT-Italian Institute of Statistics, Italy
Konstantin Todorov, Ecole Centrale Paris, France
Rafael Tolosana Calasanz, Universidad de Zaragoza, Spain
Samyr Vale, Federal University of Maranhão - UFMA - Brazil
Iván P. Vélez-Ramírez, Phidelix Technologies, Puerto Rico
Sebastian Ventura, University of Cordoba, Spain
Michael von Riegen, University of Hamburg, Germany
Rusen Yamacli, Anadolu University, Turkey
Beytullah Yildiz, Presidency of the Republic of Turkey, Turkey
Yuqing Sun, Shandong University, China
Chang-Wei Yeh, National Center for High-performance Computing - Hsinchu, Taiwan
Anastasiya Yurchyshyna, University of Geneva, Switzerland
Amelia Zafra, University of Cordoba, Spain
Fatiha Zaidi, University of Paris Sud (Orsay), France
Weiliang Zhao, Macquarie University, Australia
Martin Zimmermann, Hochschule Offenburg - Gengenbach, Germany
Christian Zirpins, University of Karlsruhe (TH), Germany
Dominik Zyskowski, Poznan University of Economics, Poland

SLAECE Track

Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong
Chi Chi Hung, Tsinghua University - Beijing, China
Eleanna Kafeza, Athens University of Economics and Business, Greece
George Pentafronimos, University of Pireaus, Greece
Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada
Yuqing Sun, Shandong University, China
Panagiotis Takis Metaxas, Wellesley College, USA

VEWAeL Track

Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Luis Borges Gouveia, University Fernando Pessoa - Porto, Portugal
Mihaela Brut, Alexandru Ioan Cuza University, Romania / IRT -Toulouse, France
Matthias Ehmann, University of Bayreuth, Germany
Jean-Pierre Gerval, ISEN Brest, France
Nael Hirzallah, Fahad Bin Sultan National University, Kingdom of Saudi Arabia
Stephanie Meerkamm, University of Bayreuth, Germany
José Raúl Romero, Universidad de Cordoba/Campus de Rabanales, Spain
Sandrine Sarre, Public Research Centre Henri Tudor - Kirchberg, Luxembourg
Antonio Sarasa-Cabezuelo, Complutense University of Madrid, Spain
Didier Sebastien, University of Reunion Island, France
José-Luis Sierra-Rodríguez, Complutense University of Madrid, Spain
Andre Luis Silva do Santos, CEFET-MA, Brazil

SERCOMP Track

Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Gustavo González-Sánchez, Mediapro Research, Spain
Hai Jin, Huazhong University of Science and Technology - Wuhan, China
Federica Paganelli, National Interuniversity Consortium for Telecommunications (CNIT), Italy
Matteo Palmonari, University of Milan - Bicocca, Milan, Italy
Omair Shafiq, University of Calgary, Canada
Ming Xue Wang, Dublin City University, Ireland

ECCSS Workshop Chair

Zaigham Mahmood, University of Derby, UK

ECCSS Workshop Co-Chair

Zhengxu Zhao, Shijiazhuang Tiedao University, China

ECCSS 2011 Technical Program Committee

Nahed Amin Azab, Regional IT Institute, Cairo, Egypt

Harjinder Singh Lalli, University of Derby, UK

Zaigham Mahmood, University of Derby, UK

Sanjay Misra, Federal University of Technology - Minna, Nigeria

Gopalakrishnan Nair, DS Institutions - Bangalore, India

Doug Thomson, RMIT University - Melbourne, Australia

Muthu Ramachandran, Leeds Metropolitan University, UK

Pascal Ravesteyn, HU University of Applied Sciences - Amsterdam, The Netherlands

Saqib Saeed, University of Siegen, Germany

Arshad Ali Shahid, National University of Computer and Emerging Science (FAST) - Islamabad, Pakistan

Zhengxu Zhao, Shijiazhuang Tiedao University, China

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission to reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Service Discovery in Ubiquitous Computing Environments <i>Luis Javier Suarez Meza, Luis Antonio Rojas Potosi, Juan Carlos Corrales, and Luke Albert Steller</i>	1
Measuring Service Cohesion Using Latent Semantic Indexing <i>Ali Kazemi, Ali Rostampour, Fereidoon Shams, Pooyan Jamshidi, and Ali Nasirzadeh Azizkandi</i>	10
Distribution and Self-Adaptation of a Framework for Dynamic Adaptation of Services <i>Francoise Andre, Erwan Daubert, and Guillaume Gauvrit</i>	16
Minimizing Human Interaction Time in Workflows <i>Christian Hiesinger, Daniel Fischer, Stefan Foll, Klaus Herrmann, and Kurt Rothermel</i>	22
Adaptive Business Process Modeling in the Internet of Services (ABIS) <i>Monika Weidmann, Falko Koetter, Maximilien Kintz, Daniel Schleicher, and Ralph Mietzner</i>	29
Context Factors for Situational Service Identification Methods <i>Rene Borner, Matthias Goeken, Thomas Kohlborn, and Axel Korthaus</i>	35
Contract-Performing Circumstance-Driven Self-Adaptation and Self-Evolution for Service Cooperation <i>Ji Gao and Hexin Lv</i>	43
Collective Service Intelligence Management in Mobiquitous Systems <i>Evgeniya Ishkina</i>	51
Service Network Modeling and Performance Analysis <i>Manolis Voskakis, Christos Nikolaou, Willem-Jan van den Heuvel, and Marina Bitsaki</i>	58
Model-Driven Dynamic Service Delivery in Mobility and Ambient Environment <i>Soumia Kessal, Noemie Simoni, Xiaofei Xiong, and Chunyang Yin</i>	64
Mechanism Design for Designing Annotation Tools <i>Roberta Cuel, Oksana Tokarchuk, and Marco Zamarian</i>	73
Performance Evaluation of Dynamic Web Service Selection <i>Miroslav Zivkovic, Hans van den Berg, Hendrik Meeuwissen, and Bart Gijsen</i>	79
The Strategic Role of IT: An Empirical Study of its Impact on IT Performance in Manufacturing SMEs <i>Louis Raymond, Anne-Marie Croteau, and Francois Bergeron</i>	89
Developing the Mobile Service Applications of a Micropayment Platform(MPP): the Perspective of Actor-	98

Network Theory <i>Jen Wel Chen, Hsiao-Chi Wu, and Ching-Cha Hsieh</i>	
Decision Method of Training Data for Web Prefetching <i>Zhijie Ban and Feilong Bao</i>	104
Towards evolvable Control Modules in an industrial production process <i>Dirk van der Linden, Herwig Mannaert, and Jan De Laet</i>	112
CincoSecurity: A Reusable Security Module Based on Fine Grained Roles and Security Profiles for Java EE Applications <i>Maria Consuelo Franky and Victor Manuel Toro C.</i>	118
A Personalized Recommender System Model Using Colour-impression-based Image Retrieval and Ranking Method <i>Ana Sasa, Yasushi Kiyoki, Shuichi Kurabayashi, Xing Chen, and Marjan Krisper</i>	124
Sharing Emotional Information Using A Three Layer Model <i>Imen Tayari Meftah, Nhan Le thanh, and Chokri Ben Amar</i>	130
An Evaluation of a Cluster-based Testbed for Peer-to-Peer Information Retrieval <i>Saloua Zammali and Khedija Arour</i>	136
Reviving the Innovative Process of Design Thinking <i>Justus Bross, Christine Noweski, and Christoph Meinel</i>	142
Promoting organisational emergence in business social networks <i>Matei Dobrescu</i>	150
Mitigating Risk in Web-Based Social Network Service Selection: Follow the Leader <i>Jebrin Al-Sharawneh, Mary-Anne Williams, Xun Wang, and David Goldbaum</i>	156
Development Problems in XML Algebraic Parsing Process <i>Adriana Georgieva and Bozhidar Georgiev</i>	165
A Privacy Policy Framework for Service Aggregation with P3P <i>Liju Dong, Yi Mu, Willy Susilo, Peishun Wang, and Jun Yan</i>	171
Dynamic music lessons on a collaborative score annotation platform <i>Veronique Sebastien, Didier Sebastien, and Noel Conruyt</i>	178
Towards a Unified User Profiling Scheme for Distributed Large Sporting Events' Environments <i>Uko Asangansi and Stefan Poslad</i>	184

Coordination based Distributed Authorization for Business Processes in Service Oriented Architectures <i>Sarath Indrakanti and Vijay Varadharajan</i>	188
Trust Metrics for Services and Service Providers <i>Zainab M. Aljazzaf, Mark Perry, and Miriam A. M. Capretz</i>	195
Towards Peer Selection in a Semantically-Enriched Service Execution Framework with QoS Specifications <i>Jun Shen, Ghassan Beydoun, Brian Henderson-Sellers, Shuai Yuan, and Graham Low</i>	201
A View-based Approach for Service-Oriented Security Architecture Specification <i>Aleksander Dikanski and Sebastian Abeck</i>	207
End-user's Service Composition in Ubiquitous Computing using Smartspace Approach <i>Muhammad Mohsin Saleemi and Johan Lilius</i>	214
Analysis and Verification of Web Services Resource Framework (WSRF) Specifications Using Timed Automata <i>Jose Antonio Mateo, Valentin Valero, Enrique Martinez, and Gregorio Diaz</i>	222
The Location-based Authentication with The Active Infrastructure <i>David Jaros, Radek Kuchta, and Radimir Vrba</i>	228
Continuous evaluation in the process of ontology development <i>Dejan Lavbic, Marjan Krisper, and Marko Bajec</i>	231
A Community Cloud: Archive Retrieval in Multiple Language Services <i>Wei-hua Zhao, Zhan-wei Liu, and Zheng-xu Zhao</i>	237
An Architecture of Virtual Desktop Cloud: Design and Implementation <i>Zhan-wei Liu, Tong-rang Fan, and Zheng-xu Zhao</i>	243
A Study of Dragon-lab Federal Experiment Cloud and Network Contest <i>Tongrang Fan, Zhanwei Liu, and Liping Niu</i>	249
Cloud Computing: Issues in Data Mobility and Security <i>Zaigham Mahmood and Harjinder Lallie</i>	255

Service Discovery in Ubiquitous Computing Environments

Luis Javier Suarez

GIT

University of Cauca
Popayán, Colombia

ljsuarez@unicauca.edu.co

Luis Antonio Rojas

GIT

University of Cauca
Popayán, Colombia

luisrojas@unicauca.edu.co

Juan Carlos Corrales

GIT

University of Cauca
Popayán, Colombia

jcorral@unicauca.edu.co

Luke Albert Steller

Faculty of I.T.

Monash, University
Victoria, Australia

Luke.Steller@gmail.com

Abstract—Today, there is an increasing abundance of information and services available to mobile users. Many ubiquitous services retrieval architectures are based on keyword or interface matching which does not provide very accurate match results. More recently, semantic languages have been used to improve accuracy. However, this often requires the use of reasoning software which is very resource intensive. Therefore, in this paper we propose a semantic approach to service retrieval in ubiquitous computing environments, which improves accuracy over keyword / interface matching approaches but avoids the use of a semantic reasoned in order to provide improved efficiency over inference based proposals. In addition, our proposal incorporates a user profile to limit the search space and takes account of the capabilities of the requesting mobile device. Our approach also transforms BPEL service descriptions into a graph to perform atomic-level graph matching. Thus, we calculate semantic similarity between two graph nodes to provide a service ranking, so that it is possible obtain an approximate match if there is no service that exactly matches the user requirements. We have implemented our approach and provide a performance evaluation on a mobile device which clearly demonstrates that our approach is more efficient than reasoning and produces accurate match results.

Keywords-matching; context-aware discovery; ubiquitous environments; personalization

I. INTRODUCTION

The number of mobile subscribers is reaching the 3 billion mark, worldwide [1]. The vision of ubiquitous computing is the amicable integration of small devices, computing and communication capabilities with humans [2] to assist them in performing their tasks, anytime and anywhere. The goal is for this integration to be as seamless as possible, ideally unconscious to the human user. Service oriented architectures [3], are useful to support transparent integration of software applications in ubiquitous environments [4]. Service discovery is used to match the requirements of a mobile user with the capabilities of existing services available. Since ubiquitous mobile environments are extremely dynamic, this matching process must be both accurate / relevant [5] and fast / efficient [6].

Service discovery in ubiquitous environments presents both new opportunities and new challenges [7, 8]. On one hand there is an abundance of contextual information about the mobile which can enrich the service discovery process. On the other hand mobile devices used in ubiquitous

environments are typically resource constrained and cannot interact with all services.

In this paper we propose a service discovery architecture for ubiquitous environments which considers the preferences of mobile users, the resource specifications of the user's device and the delivery context to provide the flexibility to reconfigure services according to environmental changes.

Typically the Business Process Execution Language (BPEL) [9] is used as an orchestration language for services. It is used to form executable business processes which involve message exchange. The number of business processes described using BPEL on the web and at an enterprise level is increasing. Additionally, BPEL is useful for forming a composition of multiple services to meet the user's requirements when a single service alone cannot perform the required task [10]. Therefore, in our approach we propose an algorithm which matches services based on BPEL descriptions.

It is well known that semantic matching is more accurate than earlier approaches such as keyword / interface based matching [11, 12]. Therefore, in order to meet the need for accuracy, our matching algorithm evaluates semantic distance between existing services. Many semantic matching approaches utilize reasoners, however, the use of reasoners has been shown to be extremely resource intensive [3, 13-15]. Therefore, in order to support efficiency we avoid the use of reasoners. Rather, we reduce the matching process to a problem of graph matching by adapting existing algorithms [16, 17]. As such our matching algorithm translates BPEL processes into graph representations then matches these graphs using semantic distance calculations [17].

We have implemented our proposed approach and provide an evaluation on a resource constrained device which shows that our approach supports both efficient matching on a resource constrained device and effectively provides accurate results.

The remainder of this paper is structured as follows: A discussion of the current research in the field is given in Section II. We present the high-level description of our architecture and matching process in Section III. Then in Section IV we discuss our approach to transform BPEL into graphs. The overall ranking process is discussed in Section V, followed by details about how two graph nodes are compared in Section VI. In Section VII we discuss the way in which our architecture filters services based on whether they are capable of running on the user's device. We provide

details about our implementation and evaluation in Section VIII. Finally in Section IX we conclude the paper.

II. RELATED WORK

Service discovery is defined as the ability to find and use a service based on a published description of its functionality and operational parameters[18]. Service discovery can be addressed under two main approaches: syntactic and semantic discovery.

Syntactic discovery is based on interface matching techniques (e.g., UDDI, ebXML, WSDL, IDL, RMI interfaces, etc.) or keywords to search for services, requiring exact matches at the syntactic level between service descriptions and parameters employees [7, 19, 20], which can result in that equivalent services at the logical level to be discarded (e.g., two services described as *printer* and *printing* may differ syntactically but logically they are equivalent).

Thus, while the syntax is focused on defining the services from the input and output messages, types and parts of the message, semantics aims to provide information about the service functionality[19, 21]. Thus, semantics improves matching accuracy. The semantic representation of service descriptions content enable machines to understand and process their content, supporting the discovery and service dynamic integration[7]. However, semantic descriptions require reasoning applications which are resource intensive applications which will significantly increase processing time[22].

Therefore, we propose a service discovery approach for ubiquitous environments based on semantic matching without a reasoner. Our approach provides a ranked list of services which completely or partially match a user request. In addition, service retrieval process considers the preferences of mobile users, the resource specifications of the user’s device and the delivery context to provide the flexibility to reconfigure services according to environmental changes.

III. ARCHITECTURE AND MATCHING PROCESS

In this section we describe our proposed architecture to perform semantic service discovery in ubiquitous environments by considering the user request, user profile and device context. In our approach, which is named U-ServiceMatch, services and user requests are described using BPEL. Figure 1 depicts our architecture which is composed of the following modules:

- *Advertiser*: Service providers advertise their services as BPEL documents, to the *Advertiser Module*, which stores this service description into the *Service Repository*.
- *Requester*: A service requesters is a mobile user which submits a BPEL request for a service.
- *BPEL Parser*: This module transforms a BPEL service description or user request into a graph, and vice versa.

- *Device Repository*: This repository stores the resource capabilities of the requesting user’s device, including processing power, screen size, input interface, etc.
- *User Repository*: This module stores details related to the mobile user / requester including personal information about the user and previously requested / invoked services.
- *Service Discovery*: This module performs the matching of a user request to service descriptions. It contains several sub-modules including the *Service Matcher* which performs the graph matching, the *Context Matcher* which determines whether services can be displayed on the user’s device and *User Matcher* which matches user profiles.

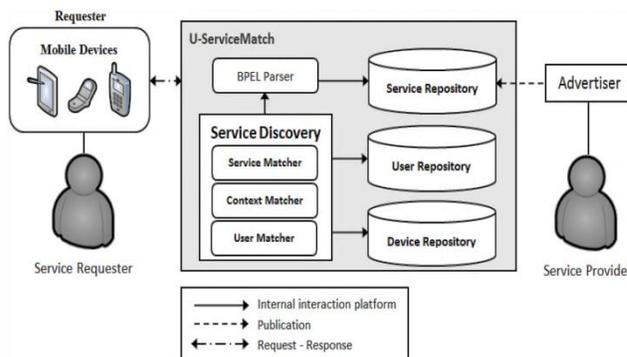


Figure 1. Architecture U-ServiceMatch

The overall module interaction is presented as an activity diagram in Figure 2. This can be described as follows.

A user submits a BPEL service description which is transformed into a graph by the *BPEL Parser*. The *Service Discovery* module then manages the matching process as follows. The user request graph is first matched (by the *Service Matcher*) with services that have been consumed in the past by the current user or other users with a similar user profile as the current user. Similar users are found by the *User Matcher* module. This step is designed to limit the search space. If a sufficiently matching service was not found, then the user request is matched by the *Service Matcher* against all other services in the *Service Repository*. Each service in the ranked list of services is checked to ensure it can be invoked / consumed by the requesting device by the *Context Matcher*. A final ranked service list is provided to the requester.

In the remainder of this paper we will discuss the following. In Section IV we will discuss the BPEL to graph transformation which is handled by the *BPEL Parser* module. In Section V we will present the overall ranking process and user profile matching handled by the *Service Discovery* module which will interact with the *User Matcher* sub-module, and the *User and Device Repositories*. In Section VI we will discuss how two graph nodes are compared by the *Service Matcher* module. In Section VII we will talk over the way in which our *Context Matcher* filters services based on whether they are capable of running on the user’s device.

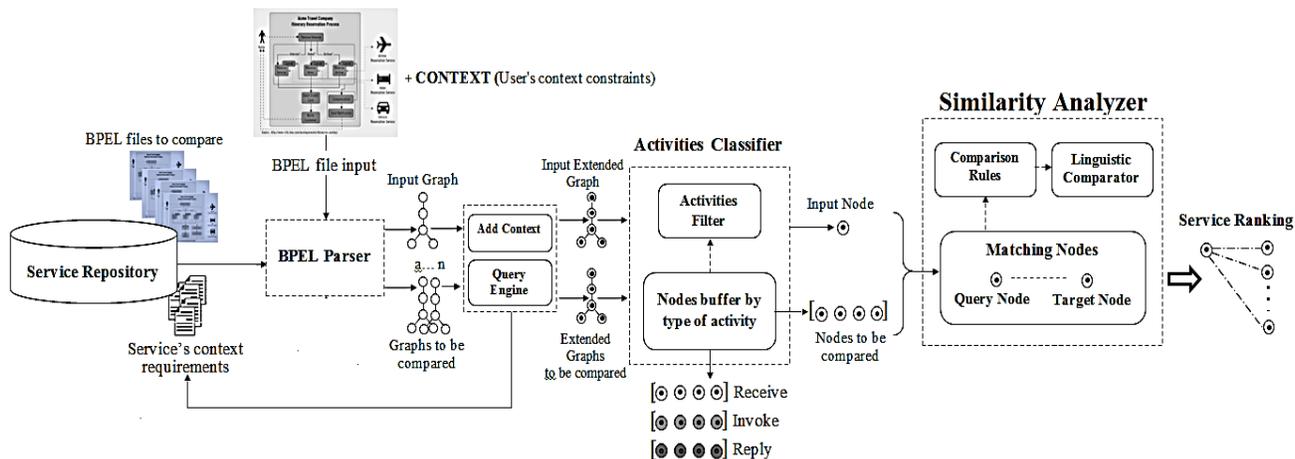


Figure 2. Matching of BPEL Basic Activities

IV. BPEL TO GRAPH TRANSFORMATION

In this work the available services in the ubiquitous network are represented by basic activities contained in a business process, denoted by BPEL. Thus, in this section, we will discuss how BPEL processes are transformed to graphs. Similarly, the nodes of the transformed graph represent the activities of the BPEL process.

Transforming BPEL to Graph: Graphs are a general and powerful data structure for representing objects and concepts. Thus, in this section will present the equivalence between a BPEL description and a formal representation of Graphs.

A graph G , in its basic form, is a pair $G = (N, E)$ where N is a non-empty finite set of elements called *nodes* (also called vertices or points) such that $N = \{n_1, \dots, n_m\}$. E is a multi-set of pairs (n_i, n_k) is not ordered distinct elements of N called *edges*, such that $E \subset N \times N$. N and E are distinct, such that $N \cap E = \emptyset$. When all the edges have directions, and therefore (n_i, n_k) and (n_k, n_i) can be distinguished, the graph is directed. Thus, a *directed graph* or *digraph* $G = (N, E)$ consists of a set N of nodes and a set E of edges, which are ordered pairs of elements of N .

The *BPEL Parser* module transforms a BPEL behavior model into a process graph. A process graph has at least one start node and can have multiple end nodes. The graph can have two kind of nodes: (1) regular nodes representing BPEL activities; and (2) BPEL connectors representing split and join rules of type XOR or AND. Nodes are connected via edges which may have an optional guard. Guards are conditions that can evaluate to *true* or *false*.

We used the flattening strategy presented in [23] to transform a BPEL document to a process graph. The general idea is to map structured activities to respective process graph fragments, Figure 3. The algorithm traverses the nested structure of BPEL control flow in a top-down manner

and recursively applies a transformation procedure to each type of structured activity.

A BPEL *basic* activity is transformed into a graph node n . The BPEL *sequence* is transformed by connecting all nested activities with graph edges; each sub-activity is then transformed recursively. For the BPEL *while* activity, a loop is created between an *XOR join* and a *BPEL XOR split*, the condition is added to the edge. The graph representation of BPEL *switch* consists of a block of alternative branches between a *BPEL XOR split* and a *BPEL XOR join*. The branching conditions are each associated with an edge. The BPEL *flow* is transformed to a block of parallel branches starting with a *BPEL AND split* and synchronized with a *BPEL AND join*.

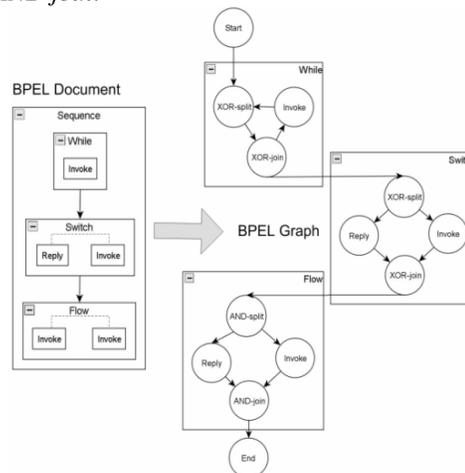


Figure 3. Correspondence between BPEL elements and Graph elements

The graph nodes n that represent BPEL activities have attributes which reflect the respective activity. These are defined as ActivityType $AT(n)$, Operation $Op(n)$, PortType $PT(n)$ and PartnerLink $PL(n)$. $AT(n)$ may contain one of the following values $Invoke^{syn}$, $Invoke^{asyn}$, *Receive* or *Reply*. The graph nodes n that represent BPEL connectors have two attributes defined as: *ConnectorType*(n) and *ActivityType*(n).

ConnectorType(n) may contain one of the following values: *AND-split*, *AND-join*, *XOR-split* or *XOR-join*. *ActivityType(n)* is the BPEL structured activity from which the node was derived during transformation. Figure 3 shows the correspondence between BPEL constructs and graph elements.

V. USER PROFILE MATCHING AND SERVICE RANKING

To produce a ranked set of services the mobile user’s service request node n_i must be matched against each service node n_j contained within a set S of potential services. There may be many potential services in the *Service Repository*. Therefore, we first check if any user has performed the same request previously, and if so obtain a ranked service list from the cache. If the request is not in the cache, the matching algorithm matches the user request against those services which have been invoked previously by the same user or a different user which has a similar user profile as the current user. If a valid service has still not been found, then the remaining services in the *Service Repository* are compared against the request.

This process is the focus of this section. First we will describe the structure of our user profiles then secondly we will describe the matching algorithm which provides a ranked list of services.

A. User Profile Structure

The structure of our user profiles is based on [24]. These profiles comprise domain of interest and personal data as shown in Figure 4. In this paper, we present a proof of concept which takes a few of these characteristics into consideration. In future work, we will expand the contextual attributes which are taken into consideration to provide a broader matching of user profile similarity.

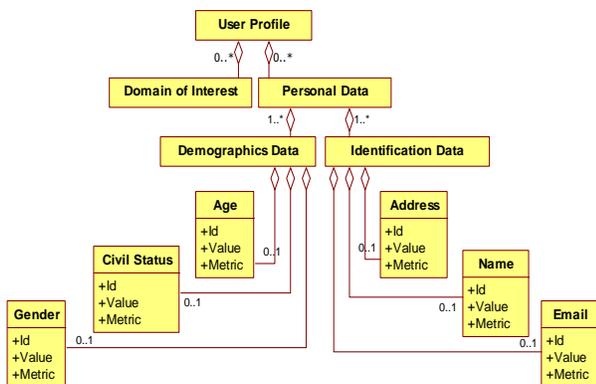


Figure 4. Meta-Model of User Profile

Several studies use different methods for collecting and handling domain of interest information, depending on the application: Web mining [25], clustering [26], Application logs [27], etc. Each of these mechanisms generates a set of parameters and their possible values for a given domain of interest. The definition of these parameters and values are not established in this work, due to the high level of analysis and decoupling to a specific field.

Personal data falls into two categories: data identification and demographics. The user profile meta-model, is stored in the *User Repository*. In our system, we compare a user profile with other profiles in order to establish a set of similar user profiles. We assume that users with similar profiles will request similar services [28]. Thus, we suggest services to a user if these have been requested or consumed by a similar user in order to reduce the search space for potential services to compare against the user request. To realize this goal, we will propose the matching process in the next subsection.

B. Rank Services

Algorithm 1 defines the algorithm for obtaining a ranked set of services which match the user request. This algorithm makes use of two functions. Let the function *GetRankedServicesFromCache(n)* provide a ranked list of services from the cache for any user request n (if one exists). If the current or another user has not submitted the request n previously, then the algorithm retrieves a list of services which the current user, or other users with a similar user profile, has invoked in the past. Let *ConsumedServices(p_i)* denote a function which returns these services, where p_i is the user profile for the current user / requester.

Algorithm 1. RankServices

```

1.  INPUTS: Node  $n_q$ , UserProfile  $p$ 
2.  OUTPUT: RankedList  $RS$  /* ranked list of service nodes */
3.  BEGIN
4.  Let  $RS \leftarrow GetRankedServicesFromCache(n_q)$ 
5.  if  $RS \neq null$  then
6.    return  $RS$ 
7.  else
8.    Let  $S \leftarrow ConsumedServices(p)$  /* where  $S$  is a set of nodes  $n_k$ , such
    that  $S = \{n_1, \dots, n_p\}$  */
9.    for each  $n_k$  in  $S$  do
10.     Let  $dist \leftarrow CheckMatch(n_k, n_q)$  /*see Alg. 3, Sec. VI*/
11.     if  $dist < 1$  then
12.        $RS \leftarrow RS \cup (dist, n_k)$  /* add  $n_k$  to set  $RS$ , ordered by
        $dist$  */
13.     end if
14.   end for
15.  end if
16.  if BadSuggest( $RS$ ) then
17.     $RS \leftarrow null$ 
18.    Let  $S = LookupServiceRepository(\text{non-operational information})$  /*
    where  $S$  is a set of nodes  $n_k$ , such that  $S = \{n_1, \dots, n_p\}$  */
19.    for each  $n_k$  in  $S$  do
20.     Let  $dist \leftarrow CheckMatch(n_k, n_q)$  /*see Alg. 3 Sec. VI */
21.     if  $dist < 1$  then
22.        $RS \leftarrow RS \cup (dist, n_k)$  /* add  $n_k$  to set  $RS$ , ordered by  $dist$ 
23.     end if
24.   end for
25.  end if
26.  return  $RS$ 
27.  END
    
```

The algorithm will obtain a match result by comparing the user request node n_i against each of these previously invoked services and add these to a ranked list. The *CheckMatch(n_p, n_q)* is a function which returns a double indicating the semantic similarity / distance between the

request node n_i and a service node n_p , which will be defined in Algorithm 3, Section VI.

Assume *ConsumedServices* passes each (p_i, p_j) pair to *CheckProfileMatch* which is defined in Algorithm 2, where p_i is the user request and p_j is all user profiles in the *User Repository* module. Algorithm 2 compares the age, marital status, gender and all interest domain attributes, associated with the two user profiles, using the algorithm *LS*, which will be defined in Algorithm 4 in Section VI.

Let *BadSuggest(RS)* denote a function which returns true if a given ranked list of services *RS*, does not contain enough services which meet a semantic similarity threshold against the user request. This condition is set by the requesting user. In the case that a service which satisfactorily matches the user request was not found (i.e. *BadSuggest* returns true), then all other services in the *Service Repository* will be compared with the service request to produce a ranked service list. Let *LookupServiceRepository* denote a function which returns the services from the *Service Repository*.

Algorithm 2.CheckProfileMatch

```

1.  INPUTSUserProfile  $p_i$ ,UserProfile  $p_j$  /* where a  $p_k$  has attributes:
    Set InterestDomains( $p_k$ ), int Age( $p_k$ ), String Marital( $p_k$ ), String
    Gender( $p_k$ ) */
2.  OUTPUT: double
3.  BEGIN
4.  Let  $m \leftarrow 0, g \leftarrow 0, \text{maxI} \leftarrow 0$ 
5.  Let  $a = 1 - \{[|\text{Age}(p_i) - \text{Age}(p_j)| / ((\text{Age}(p_i) + \text{Age}(p_j)) / 2)]\}$ 
6.  if Marital( $p_i$ ) = Marital( $p_j$ )then,  $m \leftarrow 1$ 
7.  if Gender( $p_i$ ) = Gender( $p_j$ )then,  $g \leftarrow 1$ 
8.  for each value  $v_a$  in InterestDomains( $p_i$ ) do
9.    for each value  $v_b$  in InterestDomains( $p_j$ ) do
10.     if LS( $v_a, v_b$ ) >  $\text{maxI}$  /* calculate similarity of  $p_i$  and  $p_j$  */ then
11.        $\text{maxI} = \text{LS}(v_a, v_b)$ 
12.     end if
13.   end for
14. end for
15. /*Let  $w(y)$  be a user assigned weight of importance where  $y$  is an
    attribute, Age( $p_i$ ), Marital( $p_i$ ), Gender( $p_i$ ) or InterestDomains( $P_i$ ), such
    that  $0 \leq w(y) \leq 1$  */
    
$$w(\text{Age}(p_i)) * a + w(\text{Marital}(p_i)) * m +$$


$$\text{return } 1 - \frac{w(\text{Gender}(p_i)) * g + w(\text{IntegerDomains}(p_i)) * \text{maxI}}{w(\text{Age}(p_i)) + w(\text{Marital}(p_i)) + w(\text{Gender}(p_i)) +$$


$$w(\text{InterestDomains}(p_i))}$$

16. END
    
```

In the next section we will define the *CheckMatch* function which calculates the semantic similarity between two graph nodes.

VI. ATOMIC-LEVEL GRAPH MATCHING

Matching the user request to a potential service involves the matching of two BPEL activities (as was shown in Figure 2). Let the request / query graph be denoted as G_Q and a service / target graph as G_T . Before running the matching algorithm for the nodes (n_i, n_j) where $n_i \in G_Q$ and $n_j \in G_T$, we organize / filter nodes (n_i, n_j) according to their BPEL activity type (this is completed by the *Activities Classifier* action in Figure 2). Therefore, only the nodes that

belong to the same activity type in G_Q and G_T , respectively, are compared.

The organized nodes are then compared for matching (this is completed by the *Similarity Analyzer* module shown in Figure 2). A pair of nodes (n_i, n_j) are compared by considering their semantic distance which is outlined in Algorithm 3. This algorithm also makes use of Algorithm 4 which determines the linguistic similarity between two nodes and returns a value between 1 and 0, where 1 denotes a complete match.

Algorithm 3 starts by giving priority to comparison of the operation attribute. If the two operation attributes are similar it continuing with the calculation of the similarity of other parameters (i.e. port type and partner link) to estimate the semantic distance between the two activities. In the algorithm, let $w(Op(n_i))$, or $w(PT(n_i))$, $w(PL(n_i))$, denote user specified weights of importance associated with $Op(n_i)$, $PT(n_i)$, $PL(n_i)$ in the user request, respectively.

Algorithm 3.CheckMatch

```

1.  INPUTS: Node  $n_i$ , Node  $n_j$  /* where  $n_i$  is a request node and  $n_j$  is a
    service node and a node  $n_p$  has attributes such that Op( $n_p$ ), PT( $n_p$ ),
    PL( $n_p$ ), AT( $n_p$ ) as defined in Section IV */
2.  OUTPUT: double
3.  BEGIN
4.   $OPS \leftarrow LS(Op(n_i), Op(n_j))$  /* Operation Similarity (see Alg. 4) */
5.  if  $OPS = 0$  (different Operations) then
6.    return 1
7.  else
8.  Let  $PTS \leftarrow LS(PT(n_i), PT(n_j))$  /*PortType Similarity (see Alg. 4) */
9.  Let  $PLS \leftarrow LS(PL(n_i), PL(n_j))$  /* PartnerLink Similarity (see Alg. 4) */
10. /* $w(z)$  is a weight of importance associated with an attribute  $z$  in the
    user request, such that  $z = Op(n_i)$ , or  $z = PT(n_i)$ , or  $z = PL(n_i)$ , where  $0 \leq$ 
 $w(z) \leq 1$  */
    
$$\text{Let } dist \leftarrow 1 - \frac{w(Op(n_i)) * OPS + w(PT(n_i)) * PTS + w(PL(n_i)) * PLS}{w(Op(n_i)) + w(PT(n_i)) + w(PL(n_i))}$$

11. Return  $dist$ 
12. end if
13. END
    
```

The *LS* function is defined in Algorithm 4 and is used to calculate the linguistic similarity of the values associated with the same attribute of two separate graph nodes n_i and n_j (e.g., the value of $Op(n_i)$ compared to the value of $Op(n_j)$).

Algorithm 4.LS /* LinguisticSimilarity */

```

1.  INPUTS: String  $v_i$ , String  $v_j$ 
2.  OUTPUT: double
3.  BEGIN
4.  
$$LS = \begin{cases} 1 & \text{if } (m_1=1 \vee m_2=1 \vee m_3=1) \\ m_2 & \text{if } (0 < m_2 < 1 \wedge m_1=m_3=0) \\ 0 & \text{if } (m_1=m_2=m_3=0) \\ \frac{m_1+m_2+m_3}{3} & \text{if } (m_1, m_2, m_3 \in (0,1)) \end{cases}$$

    where  $m_1 \leftarrow N\text{Gram}(v_i, v_j)$ ,  $m_2 = \text{CheckSynonym}(v_i, v_j)$ ,
 $m_3 = \text{CheckAbbreviation}(v_i, v_j)$  /* see [29] */
5.  return  $LS$ 
6.  END
    
```

In this algorithm, let *NGram*, *CheckAbbreviation* and *CheckSynonym* denote measures which are defined in [29].

*N*Gram algorithm estimates the similarity according to a common number of *q*-grams (a *q*-gram in this context refers to a sequence of letters, *q* letters long, from a given word) between the tags. *CheckSynonym* algorithm use WordNet [30] linguistic dictionary to identify synonyms, It groups English words into sets of synonyms called synsets. Synsets are interlinked by means of conceptual-semantic and lexical relations. The *CheckAbbreviation* algorithm uses a dictionary of abbreviations appropriate to the application domain. If all algorithms give a value of 1, then there is an exact match between the tags. If all give a value of 0, then there is no similarity between words. If the values produced by *CheckAbbreviation* and *Ngram* are equal to 0 and *CheckSynonym* value is between 0 and 1, the total value of the similarity is equal to *CheckSynonym*. Finally, if all three algorithms yield a value between 0 and 1, the linguistic similarity is the average of the three.

VII. CONTEXT MANAGEMENT

Since mobile users carry their device with them throughout their daily travels, there is an abundance of contextual data available which can be fed into the service matching process to provide more accurate search results [22, 31]. Our architecture captures the resource capabilities of the requesting user’s device and the resource requirements for each service. The user’s device capabilities are stored in our *Device Repository* and the service requirements of each service are stored in the *Service Repository*. After the matching process defined in the previous sections of this paper, each service in the ranked list are checked to ensure they will function on the user’s device. In the remainder of this section we will describe the structure of user context followed by the use of this information in the service ranking process.

A. User Context Structure

We capture user context characteristics such as processing power, modes of presentation, input interfaces, connectivity, etc. According to [24] context constraints, are defined as any information that could be used to characterize an entity, where an entity can be a person or object that is considered relevant to the interaction between user and an application. We propose three dimensions for defining a meta-model of user’s context:

- a) *Spatial Dimension*: contains all the parameters that are associated with geographical and spatial information of the user;
- b) *Temporal Dimension*: contains the date and time of when a service is invoked;
- c) *Device Data Dimension*: contains information related to the user’s mobile device such as installed software, operating system, processing power, available memory, etc. We capture this content using a CC/PP profile [32].

These dimensions are illustrated in Figure 5.

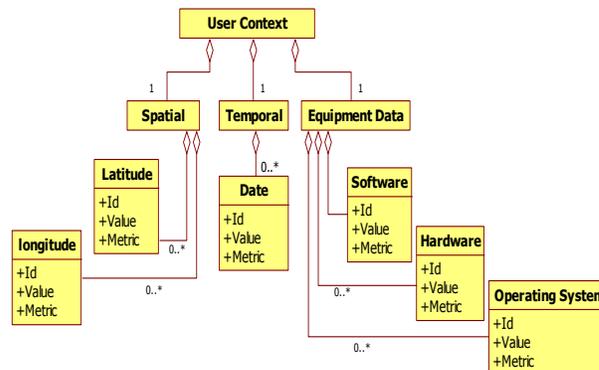


Figure 5. Meta-Model of User’s Context

The *Service Repository*, supported by the work presented in [33], stores BPEL documents and other XML files which capture the business process of services with context features. We define an XML meta-data, a model based on EMF (Eclipse Modeling Framework) for describing the restrictions specified by service providers or service developers.

The *CheckDeliveryContext* function, defined in Algorithm 5, obtains user’s context and the requirements of a particular service.

Algorithm 5. CheckDeliveryContext

```

1.  INPUTS: listRankedServices
2.  OUTPUT: Set rankedFilteredServices
3.  BEGIN
    DeviceProfile deviceProfile ← LookupDeviceProfile()
4.  for each nj in listRankedServices do
5.      EmfContext serviceContext ← LookupFeatures.Context(nj)
6.      for each ck in serviceContext do
7.          if ck ∈ deviceProfile/* requirement supported */ then
8.              rankedFilteredServices ← rankedFilteredServices ∪ nj
9.          break for
10.         end if
11.     end for
12. end for
13. return rankedFilteredServices
14. END
    
```

Algorithm 5 takes set of ranked services obtained during the matching phase, and checks each service to see whether it meets the requirements of the service context retrieved from the *Service Repository*. Let *LookupDeviceProfile* be a function which returns the device profile for the current device (i.e., from the *Device Repository*). Let *LookupFeatures.Context(n_j)* return the context requirements for a service *n_j* (i.e., from the *Service Repository*). In the case that the current user’s device can support the current service it’s added to the set which is returned, otherwise, it is discarded.

VIII. IMPLEMENTATION AND EXPERIMENTATION

This section presents the implementation and experimental study of our proposed service matching scheme for ubiquitous computing environments. Our prototype was implemented in Java. Our experiments were

completed on the following machines / devices. The server application was running on a Pentium 4, 2.30GHz processor, 1,028 MB of RAM under the OS Linux Ubuntu. We performed tests using two real client devices / phones and two emulators. The specifications for each are provided in Table I.

TABLE I. TECHNICAL SPECIFICATION OF DEVICES USED IN EACH TEST.

Device	Processor	RAM	ROM	Screen Size	Operating Systems
Pocket PC DELL AXIM x51v	Intel PXA270, 520 MHz	64M B	256M B	480 X 640 Pixels.	Microsoft Windows Mobile 5.0
Nokia N93	Dual ARM 11 332 MHz	64M B	50 MB	128 X 160 Pixels.	Symbian 9.1
Nokia 6212 NFC			Series 40 5th Edition emulator SDK		
Nokia 6260			Series 40 6th Edition emulator SDK		

A. Evaluation Methodology

We evaluated our architecture to ensure that it is both efficient and accurate. We categorized response time efficiency as follows. Let r denote response time in seconds. Let response time be classified as: *Optimal* where $r \leq 0.1$, *Good* where $0.1 \leq r \leq 1$; *Acceptable* where $1 \leq r \leq 10$; and *Deficient* where $r \geq 10$ [34]. Accuracy was measured by comparing a set of expected values against the results obtained from our architecture, using the calculations of Precision, Recall and Overall [11, 35]. Precision p is a measure of whether the list of matching services returned by our approach contains any services which were not expected to match, such that $p = x/N$, where x denotes number of services which were both expected and proven to match and N denotes the number of services found to match. Recall r is a measure of whether all of the services which were expected to match are contained in list of matching services returned by our architecture, such that $r = x/n$, where n denotes the number of services which were expected to match. The overall o match result takes account of both precision and recall such that, by $o = r * (2 - 1/p)$.

In our evaluation we created and compared 30 BPEL basic activities against 144 activities stored in the *Service Repository*, resulting in 1106 pairs to evaluate. The evaluations were done by 5 experts in service discovery, resulting in 5530 comparisons. These comparisons evaluate the attribute similarity between two BPEL basic activities. The human evaluator first made a comparison between the activities, and assigned an expected score to each activity according to their similarity to each user request, using our benchmarking tool [36]. Let s denote this score, such that $0 \leq s \leq 5$ where 0 implies no similarity / match and 5 implies complete similarity / match. The expert evaluator also sets the weights $w(z)$ for each compared attribute z to determine these expected results, which are also associated with the user requests being compared against the services in the actual system (see Algorithm 3, Section VI). The values obtained during our results were calculated using the micro-averaging technique [35].

B. Results

In this section we present the results from our tests.

1) Performance Evaluation (efficiency)

Figure 6 presents the execution times of our architecture for each of the different mobile client devices.

In each test, there were 17 BPEL files published in the *Service Repository* containing 144 target nodes or basic target activities. In addition, 5 BPEL files were used to represent 5 separate user request queries, which were each compared with the 144 target nodes.

All tests completed on the mobile devices produced results in less than 1 second for up to 144 nodes, meaning the behavior was **good**. These tests also show that our approach is substantially more efficient than using semantic reasoners which are resource intensive. For instance, in other research we used ontologies BPMO (Business Process Modeling Ontology), eTOM (enhanced Telecom Operations Map) and SID (Shared Information/Data)[37] described in WSM (Web Service Modeling Language) [38] and performed an inference / matching task on the WSM2Reasoner reasoner[39] and found that a reasoning task required approximately 170ms for just one task [13]. Our approach performed 8 comparisons in this time on the real devices (which includes network latency) and over 32 comparisons using the emulator.

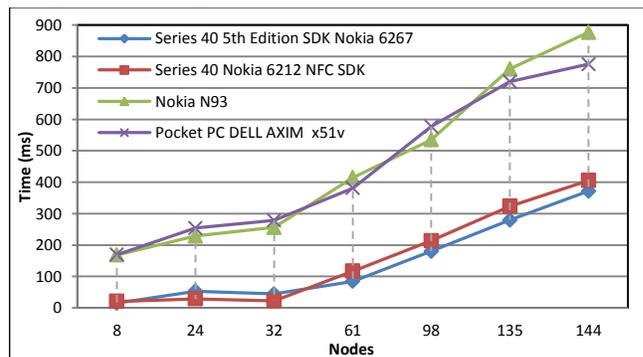


Figure 6. Recovery process performance of services on different mobile terminals.

Additionally, research shows that evaluating the control flow of BPEL documents can be exponential [17]. Our evaluation shows that our approach overcomes this problem, providing more linear results. If we extrapolate the average response time for the two real devices (i.e. Nokia and Pocket PC) presented in Figure 6 linearly[34], we can say that our architecture will have the following behavior: **Good**: when the number of graph node comparisons are less than 374.3. **Acceptable**: when the number of graph node comparisons are greater than 374.3 and less than 4145.8. **Deficient**: when the number of graph node comparisons completed are greater than 4145.8.

2) *Quality Test Results (efficacy)*: in the following we present a simulation of the service matching process on a Nokia 6260 Emulator.



Figure 7. Effectiveness Test Emulator Nokia 6260: (a) service request, (b) retrievesservice (c) service selection.

In Figure 7(a) we provide an option to select one of the 5 user request queries to compare against the available services. As shown in Figure 7(b) the user receives a listing of services which are semantically similar to the user request which was selected. In Figure 8(c) the user selects the most appropriate service from the ranked list of semantically similar services.

In Figure 8, we present the precision, recall and overall match results for our tests. A precision, recall or overall match results of 1 means that the results obtained from our architecture were equivalent to the expected results. A result of 0 means that none of the expected results were obtained.

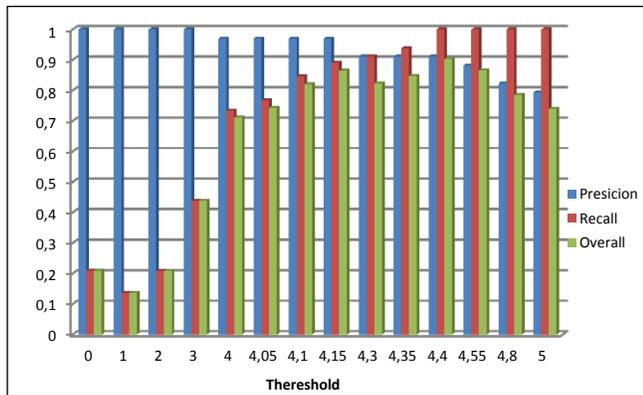


Figure 8. Quality of results produced by the U-ServiceMatch Platform

The x-axis on the graph indicates the expected similarity value s defined earlier in this section. Each bar shows an average of the precision/recall/overall results returned by U-ServiceMatch for all services with the same expected result s . We observe that services which had an expected match result of $s=4.4$ had the best precision, recall and overall match results (i.e., at least 0.9 for each). We observed that while precision was high in all tests, a recall level above 0.7 was only achieved when the threshold value was $s=4$ or above. The results also show that our approach effectively supports approximate matching of a service description with a request, when an exact match does not exist.

IX. CONCLUSION AND FUTURE WORK

In this paper we propose, develop and implement a service discovery architecture for ubiquitous computing environments. Our approach transforms BPEL user request and service descriptions into graphs which are semantically

compared to produce a ranked list of services. We also limit the search space of potential services by initially matching of the user’s request with those services which have been invoked previously by the current or other user with similar interests. Additionally, our approach filters the services which cannot be consumed on the user’s device by comparing the user’s device capabilities with the requirements of the service.

We have implemented our system as a prototype and presented an evaluation which assesses both the efficiency and accuracy of our approach. The evaluation shows that our approach is more efficient than using semantic reasoners providing **good** efficiency, performing 144 comparisons in under 1 second. We hypothesize that our approach provides acceptable efficiency for up to 4145.8 node comparisons, where acceptable implies a result was obtained within 10 seconds. U-ServiceMatch also provided extremely accurate results in terms of precision, achieving a result of 0.78-1. In terms of recall, a result of 0.7 or above was achieved with a semantic similarity threshold of 4 or above.

The next step of this work is to study and define new features that extend the user description in a ubiquitous environment. Additionally, we wish to implement a system of service registry, to reduce the search space where the *Service Repository* of considerable size in order to further improve efficiency.

REFERENCES

- [1] J. Veijalainen, "Mobile ontologies: Concept, development, usage, and business potential," *International Journal on Semantic Web and Information Systems, Special Issue on Mobile Services and Ontologies*, vol. 4, pp. 20–34 2008.
- [2] F. Almenárez, "Arquitectura de Seguridad para Entornos de Computación Ubicua Abiertos y Dinámicos," Tesis Doctoral. Departamento de Ingeniería Telemática, Escuela Politécnica Superior, 2005.
- [3] J. Zoric, N. Gjermundshaug, and S. Alapnes, "Service mobility a challenge for semantic support," presented at the 16th IST Mobile and Wireless Communications Summit, IEEE, Budapest, Hungary, pp.1-7, 2007.
- [4] C. Xiaosu and L. Jian, "Build mobile services on service oriented structure," in *IEEE International Conference on Wireless Communications, Networking and Mobile Computing*, pp. 1472–1476, 2005.
- [5] B. Kargin and N. Basoglu "Factors affecting the adoption of mobile services," *Portland International Center for Management of Engineering and Technology, IEEE*, pp. 2993–3001, 2007.
- [6] V. Roto and A. Oulasvirta., "Need for non-visual feedback with long response times in mobile hci," *International World Wide Web Conference Committee (IW3C2)*, Chiba, Japan, pp. 775-781, 2005.
- [7] S. Ben Mokhtar, *Semantic Middleware for Service-Oriented Pervasive Computing*. Tesis Doctoral, 2007.
- [8] M. Sellami, S. Tata, and B. Defude, "Service Discovery in Ubiquitous Environments: Approaches and Requirement for Context-Awareness," ed Milan, Italy: BPM Workshops, pp. 516-522, 2009.
- [9] T. Andrews, *et al.*, (2003, 05 05). Business Process Execution Language Version 1.1. the BPEL4WS Specification. Available: <http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel/ws-bpel.pdf>
- [10] V. Hermida, O. Caicedo, J.C. Corrales, D. Grigori, and M. Bouzeghoub, "Service Composition Platform for Ubiquitous Environments Based on Service and Context Matchmaking", In the

- 4th Colombian Congress of Computer, Bucaramanga, Colombia, 2009.
- [11] A. Bernstein and M. Klein, "Discovering services: Towards highprecision service retrieval," *International Workshop on Web Services, EBusiness, and the Semantic Web (CAiSE '02)*, Springer-Verlag, Toronto, Canada, vol. 2512, pp. 260 – 275, 2002.
- [12] W. Abramowicz, K. Haniewicz, M. Kaczmarek, and D. Zyskowski, "E-marketplace for semantic web services," *6th International Conference on Service-Oriented Computing (ICSOC '08)*, Springer-Verlag, Sydney, Australia, vol. 5364, pp. 271 – 285, 2008.
- [13] L. Ordoñez, A. Bastidas, C. Figueroa, and J.C. Corrales, "Task Semantic Comparison between the Telecommunications Business Processes," in *The 5th National Seminar on Emerging Technologies in Telecommunications and Telematics - TET*, Popayán, Colombia, 2010, pp. 26-31.
- [14] V. Zacharias, et al., "Mind the web," in *1st Workshop on New forms of Reasoning for the Semantic Web: Scalable, Tolerant and Dynamic in-conjunction with International Semantic Web Conference (ISWC '07) and Asian Semantic Web Conference (ASWC '07)*, vol. 291, CEUR-WS.org, Busan, Korea, 2007. Available: <http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-291/paper08.pdf>.
- [15] U. K. a. B. Köning-Rics, "Dynamic Binding for BPEL Processes - a Lightweight Approach to Integrate Semantics into Web Services," in *ICSQC*, 2007, pp. 116-127.
- [16] H. Almohamed, "A linear programming approach for the weighted graph matching problem," presented at the IEEE Trans. PAMI 15, 1993.
- [17] J. C. Corrales, *Behavioral matchmaking for service retrieval*. Versailles, France: tesis presentada a la University of Versailles Saint-Quentin-en-Yvelines para optar al grado de Doctor of Philosophy in Sciences, 2008.
- [18] A. Bandara, et al., "A Semantic Approach for Service Matching in Pervasive Environments," ed: Universidad de Southampton, 2007.
- [19] J. C. Corrales, D. Grigori, M. Bouzeghoub, and J.E. Burbano, "Bematch: A platform for matchmaking service behavior models," in *the 11th International Conference on Extending database technology: Advances in database technology (EDBT'08)*, pp. 695-699, 2008, doi: 10.1145/1353343.1353428.
- [20] W. Kokash, et al., "Leveraging web services discovery with customizable hybrid matching," presented at the In Proc. of ICSOC, 2006, pp.522-528.
- [21] E. Stroulia and Y. Wang, "Structural and semantic matching for assessing web-service similarity," presented at the Int. J. Cooperative Inf. Syst., 2005, pp. 407–438.
- [22] L. Steller and S. Krishnaswamy, "Efficient Mobile Reasoning for Pervasive Discovery," in *Proceedings of the 2009 ACM symposium on Applied Computing (SAC 2009)*, pp. 1247-1251, 2009.
- [23] J. Mendling, and J. Ziemann, "Transformation of BPEL Processes to EPCs", EPK 2005, Hamburg, Germany, vol. 167, December 2005, pp. 41-53.
- [24] E. Guerrero, J.C. Corrales, and R. Ruggia, "Service Selection based on Profile Context and QoS Metamodels", in *The 5th Conference of the Euro-American Association on Telematics and Information Systems (EATIS'10)*, 2010. ISBN 978-958-44-7280-9, in press.
- [25] S.P. Tocarruncho, F.A. Aponte, and A. Tocarruncho, "Extracción de Perfiles Basada en Agrupamiento Genetico para Recomendación de Contenido," in *Conferencia IADIS Ibero-Americana WWW/Internet*, pp. 299-303, 2007.
- [26] M. Zhang and N. Hurley, "Novel Item Recommendation by User Profile Partitioning," in *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, Milan, Italy, 2009, pp. 508-515.
- [27] S. Abbar, et al., "A personalized access model: concepts and services for content delivery platforms," in *Proceedings of the 10th International Conference on Information Integration and Web-based Applications & Services*, Linz, Austria, 2008, pp. 41-47.
- [28] S. Kurkovsky, V. Zanev, and A. Kurkovsky. "SMMART, a context-aware mobile marketing application: Experiences and lessons", *Embedded and Ubiquitous Computing*, vol. 3823, Springer-Verlag, Nagasaki, Japan, 2005, pp. 141 – 150.
- [29] J.D. Grigori, J.C. Corrales, M. Bouzeghoub, and A. Gater, "Ranking BPEL Processes for Service Discovery," vol. 3, ed: IEEE Transactions on Services Computing, pp. 178-192., 2010.
- [30] G. Miller, "Wordnet: A lexical database for english.," in *Communications of the ACM*, vol. 38 no. 11, pp. 39–41, 1995.
- [31] C. Doukeridis and N. Loutas, M. Vazirgiannis. "A System Architecture for Context-Aware Service Discovery". International Workshop on Context for Web Services (CWS'05), Paris, France, pp. 101 - 106, 2006.
- [32] Mobile W3C Device Independence Working Group, "Composite Capability / Preference Profiles (CC/PP): Structure and Vocabularies 2.0," ed, 2006.
- [33] J. Vanhatalo, J. Koehler, and F. Leymann, "Repository for business processes and arbitrary associated metadata," presented at the BPM Demo Session at the Fourth International Conference on Business Process Management, Viena. Austria, 2006, pp. 25–31.
- [34] S. Joines, R. Willenborg, and K. Hygh, "Performance Analysis for Java Websites," ed: Addison-Wesley, ISBN-13: 978-0201844542, 2002.
- [35] D. Lewis, "Representation and learning in information retrieval," in *Ph.D. Thesis*, ed University of Massachusetts: Department of Computer and Information Science, 1992.
- [36] L.J. Suarez, L.A. Rojas, J.C. Corrales, and O.M. Caicedo,(2010) Be4SeD: Benchmarking for evaluation of Service discovery techniques. *Revista de Ingeniería y Competitividad, Universidad del Valle*, in press.
- [37] A. Duke, J. Davies, M. Richardson, and N. Kings, "A Semantic Service Orientated Architecture for the Telecommunications Industry ", *INTELCOM*, vol. 3283, 2004, pp. 236-245, doi: 10.1007/978-3-540-30179-0_21.
- [38] WSMML. *Web Service Modeling Language*. Available: <http://www.wsmo.org/wsmml/index.html>
- [39] WSMML2Reasoner. *WSMML2Reasoner framework*. Available: <http://tools.sti-innsbruck.at/wsmml2reasoner/>

Measuring Service Cohesion Using Latent Semantic Indexing

Ali Kazemi, Ali Rostampour, Fereidoon Shams, Pooyan Jamshidi, Ali Nasirzadeh Azizkandi

Automated Software Engineering Research Group

Electrical and Computer Engineering Faculty, Shahid Beheshti University GC,

Tehran, Iran

E-mail : {Ali.Kazemi,A.Rostampour}@mail.sbu.ac.ir

{F_Shams, P_Jamshidi}@sbu.ac.ir, Ali_Nasirzadeh@tabrizu.ac.ir

Abstract—As low coupling, high cohesion is a service-oriented design and development principle that should be kept in mind during all stages. High cohesion increases the clarity and ease of comprehension of the design that simplifies maintenance and achieves service granularity at a fairly reasonable level. However, unlike coupling that only measures the degree of structural and behavioral dependency to the other services, cohesion metrics need to evaluate the degree of semantic relationships between operations within a service in order to measure functional relatedness. Latent semantic indexing (LSI) is one of the techniques in the field of information retrieval which is widely used to measure the degree of semantic relatedness between a document and a given query and also used to measure the cohesion of a text. In this paper, we propose an approach to automatically measure the strength of conceptual cohesion of a service based on LSI technique. Finally, it has been evaluated theoretically based on a set of cohesion principles.

Keywords—Service cohesion; Latent semantic indexing; Software metric.

I. INTRODUCTION

Service-oriented architecture (SOA) is a promising solution to build enterprise application programs which supports processes and functions as a set of well-defined services [1], [2]. Simply, a service is defined as a set of related operations. Thus, it is a logic encapsulated by individuals which supposed to be reusable [3]. Considering design standards, which causes services to be potentially reusable, the chance of a service to be able to accommodate future requirements with the least development effort increases [4]. Therefore, reusability is an important quality attribute which must be measured to satisfy an important need in SOA, the need for independent services to deliver a reusable functionality [3], [4]. One of the design attributes which has a great impact on reusability of a particular service, is cohesion, so that higher cohesion significantly increases service reusability [5], [6]. Also in [7], there is an elaborate discussion about the impact of cohesion on maintainability. The higher the cohesion of a service, the easier the test and analysis and higher cohesion will improve system stability and changeability [8] and, consequently maintainability of a system will be improved.

Because of inherently conceptual nature, cohesion is one of the most complicated and difficult structural attribute of a

class, component, or a service from quantifying point of view [8]. This quality attribute can be measured based on conceptual relatedness degree of operations which are exposed in service interface. According to a cohesion category proposed in [6], [7], conceptual cohesion is considered as the strongest type of cohesion. However, this type of service cohesion cannot be easily measured using the previous traditional metrics due to additional level of abstraction and highlighted characteristics of service interfaces in comparison with procedural and object-oriented paradigms [6].

Using the concept of latent semantic indexing (LSI), we will evaluate the conceptual relatedness degree of operations existing in a service. For the first time, LSI was used in information retrieval techniques [9], [10]. One of the applications of this method is measuring cohesion of a text [11]. LSI provides completely automatic approach that compares information units in order to measure conceptual relatedness. Measuring conceptual relatedness degree of units relies on a powerful mathematical method called Singular Value Decomposition (SVD) [10], [11]. Therefore, the objective of this research work is to propose a LSI-based approach for measuring the degree of conceptual cohesion in a service.

In order to adopt the LSI technique to measure service cohesion, we get the required information of interactions between business processes and business entities that are mostly used during service identification [6]. The SVD method is applied on a well-defined structure comprising this information. The output of this algorithm is used to quantitatively measure conceptual cohesion degree of the identified services. Utilization of semantics existing in enterprise processes is completely proportional to this inherently conceptual nature, and therefore we will have a more precise measurement of conceptual cohesion of services.

The rest of this paper is organized as follows. Section II introduces the most related works. In Section III, basic concepts of the utilized terminologies are defined. The LSI concepts and the way of adopting them are introduced in Section IV. The proposed metric and complementary example and issues are discussed in Sections V, VI and VII, respectively. The theoretical principles of the metric are

evaluated in Section VIII. Finally, the conclusion, which leads to further research, is explained in Section IX.

II. RELATED WORK

In this section, we briefly present some of the previous works on measuring cohesion in service-oriented, object-oriented and procedural paradigms. The concept of cohesion in OO and procedural paradigms has been widely discussed and examined. For example, in [12], six semantic categories of procedural cohesion namely Coincidental, Logical, Temporal, Communicational, Sequential, and Functional have been proposed. The concept of cohesion later was extended by Eder et al. [13] to cover conceptual and technical features introduced in OO paradigm. Eder et al. [13] proposed five cohesion categories (from the weakest to strongest): Separable, Multifaced, Non-delegated, Concealed and Model. Moreover, in [6], eight semantic categories of service-oriented cohesion are proposed. These categories are: Coincidental, Logical, Temporal, Communicational, External, Implementation, Sequential, and Conceptual. In [6], four categories namely Communicational, External, Implementation, and Sequential are represented as quantifiable cohesion categories. On the other hand, four categories namely Coincidental, Logical, Temporal, and Conceptual are identified in this paper as purely semantic cohesion categories. They believe that second four categories are semantic based whereas first ones are measurable without considering semantic issues. The proposed quantifiable cohesion categories have indirect impact on conceptual cohesion. A brief representation of the cohesion metrics has been shown in Table I.

TABLE I. SUMMARY OF COHESION METRICS IN THE LITERATURE

Name	Definition
<i>LCOM</i> [14]	Non-similar method pairs are counted in a class of pairs.
<i>LCOM3</i> [15]	The number of connected components in Graph is counted: Nodes are methods and edges are connections between similar methods.
<i>RLCOM</i> [16]	Number of non-similar method pairs, to The total number of method pairs ration in the class.
<i>TCC</i> [17]	Ratio of number of similar method pairs to total number of method pairs in the class.
<i>WTCoh</i> [18]	Number of used shared data entities by methods and also taking the transitive cohesion into account.
<i>SIDC</i> [7]	Number of shared parameters of the service operations divided by the total number of parameters
<i>DM IAUM</i> [19]	Number of system services divided by the total number of used messages
<i>SIDC</i> [6]	This metric is introduced to measure communication cohesion and it considers parameters and common return types.
<i>SIUC</i> [6]	This metric describes that a service is externally cohesion when all of its operation are invoked by all clients of this service.
<i>SIIC</i> [6]	This metric describes that a service has implementation cohesion when its all operations are implemented by the same implementation.
<i>TICS</i> [6]	A service is deemed to be Sequentially cohesive when all of its service operations have sequential dependencies, where a post condition/output of a given operation satisfies a precondition/input of the next operation.

It is worth to mention that in [20] and in [8], in addition to the number of shared parameters, other shared attributes such as number of service consumers, operations sequence and some more shared attributes are considered. To the best of our knowledge, there is no metric which measures the degree of relationship between operations of a service from conceptual point of view. Most proposed cohesion metrics in previous studies expect the services to have common inputs and outputs and does not consider the inter-relation of their parameters. To measure the conceptual cohesion, we require additional semantics. Therefore, we should look for methods that can measure the strength of conceptual relationship between two operations of a service by means of assets which are available in design level (processes from which services are identified) and then propose a metric for measuring conceptual cohesion of a service.

III. BASIC CONCEPTS

In this section, we present definitions of several key notions that will be utilized in this paper.

Definition 1 (Business Entity): A business entity (BE) is a dominant information entity with an associated data model and an associated behavior model in the context of a process scope [23].

Definition 2 (Elementary Business Process): An elementary business process (EBP) can be defined as $EBP = \{n, (BE_j, sr)\}$, where n is the name of elementary business process, BE_j is the jth business entity which semantically relate to corresponding EBP. $sr = \{ "C", "R", "U", "D" \}$ is the type of semantic relationship between EBP and BE_j [22].

Definition 3 (CRUD matrix): A CRUD matrix can be defined as $M = \{(EBP_i, BE_j) \ i=1 \dots \#row, j=1 \dots \#column\}$, where EBP_i is the ith EBP and BE_j is the jth BE. #row is the number of EBP and #column is the number of BE in the model [22].

Definition 4 (Conceptual Cohesion) : There is a meaningful semantic relationship between all operations of a service in terms of some identifiable domain-level concept. [6].

IV. APPLICABILITY OF LSI IN COHESION MEASUREMENT

LSI is a vector model-based technique which is applied in many information retrieval applications. In the vector model, each document is simply represented by a $A_{n \times m}$ term-document matrix, where n is the number of terms and m is the number of documents in the collection. Each cell, $a_{i,j}$, is the frequency of term t_i in the document d_j . LSI technique includes the following main steps:

1. A matrix is formed; each row of this matrix is corresponded to a term which occurs in the document. Each element (m,n) in the matrix is corresponded to number of times that term m occurs in document.
2. Local and global weighting of terms is applied to each element of the term-document.
3. SVD is used by LSI and decomposes the matrix into three other matrices: T, a term in the dimension; S, a

diagonal matrix of singular values, and D a document matrix in the dimension. The number of dimensions is considered as $t = \min(m, n)$ where m and n are the number of the terms and the number of documents in the main term-document matrix respectively. The matrix can be provided by $A = TSD^T$ where D^T is transposed of matrix D.

4. In the LSI system, the T, S and D matrices are truncated to k dimensions. Dimensional reduction reduces “noise” in the term-term matrix resulting in a richer term relationship structure that reveals latent semantics and is a crucial step in this research work.

Now we explain each one of the above steps in more details.

In the first step, the term-document matrix A is formed.

In the second step, a weight is assigned to each term in the document. There are different weighting models which are explained in [20][9]. The simplest weighting model can be obtained simply by counting number of frequency of a term in the document. In order to put the weights in the interval [0,1], the weight of each term is divided in document by tf_{max} , where tf_{max} is the maximum of term in the document.

In the third step, term-document matrix, A is taken and then is decomposed into three matrices T, S, and D using SVD. Matrices T, S and D keep the information related to terms, singular values and documents respectively.

In the fourth step, T, S, and D matrices are decreased to K domains. After dimensional reduction, the term-term matrix can be approximated using the formula: $TTS = T_K S_K (T_K S_K)^T$.

In this work, we suppose that the value which exists in location (i,j) of TTS matrix show the similarity between terms i and j in the collection. The value of K is optional, in this paper according to [20], $K=2$. Our main goal in this paper is to present cohesion metric which is able to measure the strength of conceptual similarity between operations of a service. In the following lines we explain the way of mapping above concepts to the ones which exist in SOA.

Similar to LSI, we define the BE-EBP matrix $A_{n \times m}$ of enterprise processes and business entities. Each (i,j) element in the matrix A shows the weight of i th business entity in the j th business process which is defined as the number of times that j th business process accesses the i th business entity. Each process is considered as a document. For example, the claim business process in [23] is able to access three business entities Loss Event, Claim, and Payment. The related row to this process in matrix is shown in Table II.

TABLE II. THE ASSOCIATED BE-EBP MATRIX

		Process \longleftrightarrow			
		P1			
Business Entity \updownarrow	Claim	10			
	Loss Event	3			
	Payment	5			

As shown in Table II, process P1 have accessed the Claim, Loss Event, and Payment business entities 10, 3, and 5 times respectively. The above matrix is completed for all enterprise business processes in a way that the number of its columns is equal to enterprise business processes and the number of its rows is equal to enterprise business entities. Then, three matrices are obtained using SVD. Considering $K=2$, the reduced matrix $TTS = T_2 S_2 (T_2 S_2)^T$ is formed.

TTS matrix shows the relationship between business entities. The values of the elements in TTS matrix are not normalized, and they can even be negative. Since negative values have no meaning, we substitute it by zero which means no degree of cohesion between service operations. Also to normalize values, we multiply matrix by $1/max$. Where max is the greatest value in the TTS matrix. Therefore, using the LSI concepts, we could show the existing semantic in business process in the form of TTS matrix. Finally, we use this matrix to obtain the relationship between operations of a service.

V. THE PROPOSED METRIC

The metric will be introduced in this section can be used for measuring the cohesion of a service in design time, based on the exposed operations in its interface. Note that the proposed metric is defined on an absolute scale, where a value is assigned to it in a range between 0 to 1. Value 1 shows the strongest cohesion and 0 shows lack of cohesion. Values between 0 and 1 are considered as different degrees of cohesion.

As we mentioned in Section IV, to measure a service cohesion using the proposed metric, first the BE-EBP matrix should be formed. This matrix can be formed based on those enterprise processes which services are intended to be obtained from their decomposition as defined in definition 3.

The measuring procedure has the following form. Firstly a matrix $A_{n \times m}$ is formed where m is the number of enterprise business entities and n is the number of enterprise processes. Then the number of times that each business entity i accessed by business process j , is considered as element (i,j) of matrix A. In order to obtain conceptual relatedness between business entities, we apply SVD on matrix A. Its outputs are three matrices which are shown as $A = TSD^T$.

As we discussed earlier, $TTS = T_2 S_2 (T_2 S_2)^T$ matrix shows the conceptual relatedness between business entities which are used to obtain the strength of service cohesion. For this purpose, we use a graph based approach.

Suppose that service S has a set of operations $O = \{O_1, O_2, \dots, O_m\}$. Each operation O_j of the service S accesses a set of business entities which is shown as $BE_j = \{BE_{j,1}, BE_{j,2}, \dots, BE_{j,n}\}$. For each pair of operations O_i and O_j in the service S we form a complete graph $G=(V,E)$ so that $V = BE_i \cup BE_j$.

Now, in set E, we assign a value for each edge that represents the degree of relationship between business entities, which is considered as nodes in graph G. The degree of relationship between two business entities can be measured from TTS matrix. The degree of conceptual

relatedness between two operations i and j is calculated through formula:

$$OCV(i, j) = \begin{cases} \frac{\sum_{p \in V} \sum_{q \in V} TTS_{p,q}}{|V| \times (|V|-1) / 2} & |V| > 1 \\ 1 & |V| = 1 \end{cases} \quad (1)$$

where:

- p and q are two business entities in V .
- $TTS_{p,q}$ is the degree of relationship between two business entities, BE_p and BE_q .
- $|V|$ is the cardinality of set V .
- The denominator is the number of edges in the complete graph G .

The strength of cohesion is defined as the degree of relationship between service's operations.

$$SCV(S) = \begin{cases} \frac{\sum_{i \in O} \sum_{j \in O} (OCV(i, j))}{m \times (m-1) / 2} & |m| > 1 \\ 1 & |m| = 1 \end{cases} \quad (2)$$

where:

- m is the number of operations in service S .

VI. EXAMPLE

In this section, we show how the proposed metric works using an example. To do that, we must have the enterprise processes and services which are identified using those processes. Using a real-world business process the effectiveness of the proposed metric is studied and evaluated. The sales department is studied in this scenario [22].

Using CRUD matrix is one of the ways to identify a service [22]. Table III illustrates the CRUD matrix associated to our scenario. Identified services are shown in the form of clusters with different colors (Table III).

TABLE III. THE CRUD MATRIX FOR SALES DEPARTMENT SCENARIO

EBP \ BE	customer	Credit	Account receivable note	Order	Discounts	Invoice	Shipping schedule	Draft	Inventory	Warehouse voucher
Add Customer	C	C								
Add an Account receivable note	R	U	C			R				
Check Credit	R	R			R					
Receive order	R			C						
Calculate discounts				R	R					
Check inventory				R	R				R	
Calculate price				R	R					
Add discounts				R	C					
Issue invoice	R	R		R		C				
Schedule shipping						R	C			
Issue draft						R	R	C		
Add an Item									C	
Add a warehouse voucher	R					R			U	C

The BE-EBP matrix is shown in Table IV. Since EBPs existing in the CRUD matrix access each business entities just 0 or 1 time, elements of this matrix are just 0 and 1. For example, Add Customer accesses only customer and credit

BEs, therefore there are just two ones in Add Customer column.

TABLE IV. THE BE-EBP MATRIX

EBP \ BE	Add Customer	Add an Account note	Check Credit	Receive order	Calculate discounts	Check inventory	Calculate price	Add discounts	Issue invoice	Schedule shipping	Issue draft	Add an Item	Add a warehouse voucher
Customer	1	1	1	1	0	0	0	0	1	0	0	0	1
Credit	1	1	1	0	0	0	0	0	1	0	0	0	0
Account receivable note	0	1	0	0	0	0	0	0	0	0	0	0	0
Order	0	0	1	1	1	1	1	0	1	0	0	0	0
Discounts	0	0	0	0	1	0	1	1	0	0	0	0	0
Invoice	0	1	0	0	0	0	0	0	1	1	1	0	1
Shipping schedule	0	0	0	0	0	0	0	0	0	1	1	0	0
Draft	0	0	0	0	0	0	0	0	0	0	1	0	0
Inventory	0	0	0	0	0	1	0	0	0	0	0	1	1
Warehouse voucher	0	0	0	0	0	0	0	0	0	0	0	0	1

After the matrix of BE-EBP is obtained, we apply SVD algorithm on it. To do that, MATLAB version 7.6.0.324 has been used. To obtain business process entity matrix we use this equation:

$$TTS = T_2 S_2 (T_2 S_2)^T \quad (3)$$

The resulted matrix has been shown in Table V. Also this matrix has been normalized and its negative values have been substitute with 0s.

TABLE V. THE BE-BE MATRIX AFTER DECOMPOSITION AND NORMALIZATION

BE \ BE	Customer	Credit	Account receivable	Order	Discounts	Invoice	Shipping schedule	Draft	Inventory	Warehouse voucher
Customer	0	0.99	0.27	0.94	0.15	1.00	0.19	0.10	0.33	0.21
Credit	0.99	0	0.19	0.68	0.11	0.72	0.14	0.07	0.23	0.15
Account receivable	0.27	0.19	0	0.23	0	0.28	0.08	0.04	0.06	0.06
Order	0.94	0.68	0.06	0	0.57	0.14	0	0	0.22	0.03
Discounts	0.15	0.11	0	0.57	0	0	0	0	0.03	0
Invoice	1.00	0.72	0.28	0.14	0	0	0.36	0.19	0.23	0.24
Shipping schedule	0.19	0.14	0.08	0	0	0.36	0	0.08	0.04	0.08
Draft	0.10	0.07	0.04	0	0	0.19	0.08	0	0.02	0.04
Inventory	0.33	0.23	0.06	0.22	0.03	0.23	0.04	0.02	0	0.05
Warehouse voucher	0.21	0.15	0.06	0.03	0	0.24	0.08	0.04	0.05	0

Next we show how to calculate the cohesion of a service using the proposed metric. Table III show a CRUD matrix with four identified services. First we show how to calculate the metrics for the first service which is shown by blue color.

The service has three operations which are specified by following names: Add Customer, Add an Account receivable note, Check Credit.

We have:

$$O = \{O_1, O_2, O_3\}$$

$$BE_1 = \{Customer, Credit\}$$

$$BE_2 = \{Customer, Credit, Accountreceivablenote\}$$

$$BE_3 = \{Customer, Credit\}$$

In order to obtain conceptual relatedness between operations of a service we use a graph. For operations O_1 and O_2 , graph $G = (V, E)$ has the form of Figure 1. In this graph the set V has the following form.

$$V = BE_1 \cup BE_2 = \{Customer, Credit, Accountreceivablenote\}$$

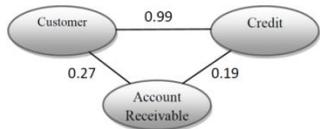


Figure 1. Business Entities Graph for Service 1

$$OCV(1,2) = \frac{0.9967 + 0.1954 + 0.2712}{3} = 0.4877$$

The results of the alternatives for the specified service S1 is shown in Table VI.

TABLE VI. OCV VALUE FOR SERVICE 1

Metric	O _i ,O _j	O ₁ ,O ₂	O ₁ ,O ₃	O ₂ ,O ₃
OCV		0.4877	0.9967	0.4877

Finally, the strength of conceptual cohesion of service is obtained.

$$SCV(S_1) = \frac{0.4877 + 0.9967 + 0.4877}{3} = 0.6573$$

Table VII shows the conceptual cohesion of four identified services on CRUD matrix of Table III.

TABLE VII. SCV VALUES FOR IDENTIFIED SERVICES

Service	The value of cohesion (SCV)
S ₁	0.6573
S ₂	0.6256
S ₃	0.2658
S ₄	0.0524

VII. DISCUSSION

The results clearly demonstrate that our proposed metric for cohesion appropriately measure conceptual cohesion of a service. Now, we analyze the values which provided by the proposed metric and the previous metrics such as SIDC [6] and TCC [17] and CCM [22]. As explained earlier, the operations of a service must be related in terms of some domain-level concepts. In other words, they must be focused on single business functionality. The analysis shows that semantics in business process are utilized properly in the proposed metric, so it evaluates service cohesion completely in conceptual point of view.

We consider two identified services in CRUD matrix (Table III), highlighted with red and blue colors, as the material for analysis. Each of these services has three operations; their operations and the resulted cohesion value, obtained by three mentioned metrics have been shown in Table VIII. Values shown in Table VIII state the relationship between two operations of each service. Consider group B1 of the first service and group R3 of the second service. For B1 and R3 groups, the SIDC and TCC give the same result value whereas these groups have different cohesion in conceptual point of view.

TABLE VIII. COHESION VALUES OBTAINED BY MENTIONED METRICS

Cluster	Group	EBP Index	Proposed Metric	SIDC	TCC
Blue	B1	1,2	0.48	0.66	0.66
Red	R3	10,11	0.21	0.66	0.66

In B1, Credit and Customer BEs have been accessed together four times (by 1, 2, 3, 9 EBP of CRUD matrix), these two entities are very related conceptually, because according to [22] two BEs are related if there is at least one shared activity in their behavioral model. In behavioral model of Credit and Customer, there are four shared activities that are processed simultaneously. Therefore, any action on one of them requires an action on the other. In other words, in this organization, whenever an operation performed on Customer, we can expect that an operation must be done on Credit entity. Thus generally, we can associate performing an action on one of them with performing an action on the other as an atomic activity (Create, Update, Read and Deletion of Credit entity and vice versa). On the other hand, existing high cohesion between service operations can be considered as a reusability predictor. This capability is provided by the proposed metric clearly. It is obvious that whenever an operation on a BE is performed with another operation on another BE frequently in enterprise processes, means that performing these two operations together has higher potential reusability. Consequently, it is better to place these two operations, which are considered as an atomic activity, in the same service. Existing of Account Receivable Note in this service (Blue Cluster) results in corruption of this service. Because this BE shares one activity in its behavioral model with behavioral model of other two entities (second EBP of CRUD matrix). Therefore, the cohesion value of 0.48 has been obtained for group B1. In R3 group, two BEs, Issue and Shipping, have been accessed together just two times (by 10,11 EBP of CRUD matrix), so we can say that these entities are less related in conceptual point of view in comparison with Credit and Customer entities. Moreover, Draft has just one shared activity with two other entities in its behavioral model (EBP11 of CRUD matrix). By conducting similar analysis, there is a lower cohesion between existing operations in R3 in comparison with B2 which our metric shows this point by obtaining cohesion value of 0.21.

VIII. EVALUATION OF THE METRIC

The proposed cohesion metric is analytically evaluated by using property-based software engineering measurement framework [24]. The metric satisfies all of the cohesion properties and therefore it can be a valid measure of cohesion from the measurement theory point of view.

Property 1: Non-negativity and Normalization are satisfied because SCV metric never becomes negative under any conditions, and its value will be in [0,1]. Normalization always let the direct and meaningful comparison between strength of services' cohesion.

Property 2: Null Value is satisfied because SCV metric gets the value if mutual relationship between all of business entities which are used by operations of a service is 0.

Property 3: Monotonicity is satisfied because by adding a related business entity to a pair EBP its overall cohesion is not decreased. In the other words, whenever we add a related business entity to a set of BEs which are accessed by a pair service operations, the cohesion between those two operations will not be decreased.

Property 4: Cohesive Modules is satisfied because by joining two unrelated service interface, the resulted cohesion will not be greater than the cohesion of original interfaces. In the other words, the strength of cohesion between operations of two unrelated services will not be greater than the strength of each service, because they access unrelated BEs.

IX. CONCLUSION AND FUTURE WORK

In this paper, using LSI technique, the strength of conceptual cohesion of a service was measured. In this technique, the business entity-business process matrix is formed using existing semantics in business processes and then by applying SVD algorithm on this matrix, the business entity-business entity matrix was resulted so that this matrix represented conceptual relationship between business entities. By adopting business entity-business entity matrix, we can measure the strength of conceptual cohesion of candidate services in the service identification phase. This quality attribute has a great impact on service reusability and maintainability but inherently conceptual nature caused it to be very difficult from quantifying point of view. Therefore measuring this important quality attribute from the conceptually point of view is very valuable. Writers' lack of access to all enterprise processes caused that the effectiveness of the cohesion measuring approach to be shown using a CRUD matrix. Although the obtained results approves the effectiveness of proposed metric well, but having all processes of a real enterprise and then using this metric in the service identification phase completely approves usefulness of this metric. Therefore, using more case studies can be considered as future work of this paper.

ACKNOWLEDGMENT

The project has been partially supported by Iran Education and Research Institute for Information and Communication Technology (formerly Iran Telecommunication Research Center (ITRC)) under contract number 4290/500 and also Shahid Beheshti University under the supervision Automated Software Engineering Research (ASER) Group, Faculty of Electrical and Computer Engineering.

REFERENCES

- [1] A. Erradi, N. Kulkarni, and P. Maheshwari, "Service Design Process for Reusable Services: Financial Services Case Study," 5th International Conference on Service Oriented Computing (ICSOC'07), 2007.
- [2] M.P. Papazoglou and W.-J. van den Heuvel, "Service-Oriented Design and Development Methodology," International Journal of Web Engineering and Technology, vol. 2, no. 4, pp.412-442, 2006.
- [3] M. P. Singh and M. N. Huhns, "Service-Oriented Computing: Semantics, Processes," Agents. John Wiley and Sons, WestSussex, England, 2005.

- [4] T. Erl, "Service-Oriented Architecture (SOA): Concepts, Technology, and Design," Prentice Hall, 2005.
- [5] J. Bansiya and C. G. Davis, "A Hierarchical Model for Object-Oriented Design Quality Assessment," IEEE Transactions on Software Engineering, 2002.
- [6] M. Perepletchikov, C. Ryan, and Z. Tari, "The Impact of Service Cohesion on the Analyzability of Service-Oriented Software," IEEE TRANSACTIONS ON SERVICES COMPUTING, vol. 3, No. 2, 2010.
- [7] M. Perepletchikov, C. Ryan, and K. Frampton, "Cohesion Metrics for Predicting Maintainability of Service-Oriented Software," Seventh International Conference on Quality Software (QSIC 2007), 2007.
- [8] L.C. Briand, J.W. Daly, and J. Wust, "A Unified Framework for Cohesion Measurement in Object-Oriented Systems," FourthInt'l Software Metrics Symp, 1977.
- [9] E. Greengrass, "Information Retrieval: A Survey 30 November 2000," 2000.
- [10] S. Dominich, The Modern Algebra of Information Retrieval, Springer, 2008.
- [11] P. W. Foltz, W. Kintsch, and T. K. Landauer, "The Measurement of Textual Coherence With Latent Semantic Analysis," Discourse Processes, 1998.
- [12] W. Stevens, G. Myers, and L. Constantine, "Structured Design," IBM Systems J., vol. 13, no. 2, pp. 115-139, 1974.
- [13] J. Eder, G. Kappel, and M. Schrefl, "Coupling and Cohesion in Object-Oriented Systems," ACM Conf. Information and Knowledge Management (CIKM), 1992.
- [14] C.K. Chidamber and S. R. Kemerer, "A Metrics Suite for Object Oriented Design," IEEE Transactions on Software Engineering, Vol. 20, 1994.
- [15] W. Li and S. Henry, "Object-Oriented metrics that predict maintainability," Journal of Systems and Software, 1993.
- [16] X. Li, Z. Liu, B. Pan, and D. Xing, "A Measurement Tool for Object Oriented Software and Measurement Experiments with It," In Proc. IWSM 2000. (Lecture Notes in Computer Science 2006, Springer-Verlag, Berlin, Heidelberg, 2001), pp.44-54.
- [17] J. M. Bieman and B-Y. Kang, "Cohesion and Reuse in an Object-Oriented System," In Proc. ACM Symposium on Software Reusability (SSR'95), 1995.
- [18] G. Gui and P.D. Scott, "Coupling and Cohesion Measures for Evaluation of Component Reusability," MSR'06, 2006.
- [19] B. Shim, S. Choue, S. Kim, and S. Park, "A Design Quality Model for Service-Oriented Architecture," 15th Asia-Pacific Software Engineering Conference, 2008.
- [20] P. Ingwersen, Information retrieval, Taylor Graham Publishing, 1992.
- [21] M. Qian, N. Zhou, Y. Zhu, and H. Wang, "Evaluating Service Identification with Design Metrics on Business Process Decomposition," IEEE International Conference on Services Computing, 2009.
- [22] A. Rostampour, A. Kazemi, F. Shams, A. Zamiri, and P. Jamshidi, "A Metric for Measuring the Degree of Entity-Centric Service Cohesion," Service oriented Computing and Application (SOCA), IEEE, 2010, in press.
- [23] S. Kumaran, R. Liu, and F. Y. Wu, "On the Duality of Information-Centric and Activity-Centric Models of Business Processes.", "Springer, CAiSE", pp. 32-47, 2008.
- [24] L. C. Briand, S. Morasca, and V. R. Basili, "Property-Based Software Engineering Measurement.", "Transactions on Software Engineering", Issue. 1, Vol. 22, 1996.
- [25] A. Rostampour, A. Kazemi, F. Shams, P. Jamshidi, and A. Nasirzadeh Azizkandi, "Measures of Structural Complexity and Service Autonomy," International Conference on Advanced Communication Technology, IEEE, 2011, in press.

Distribution and Self-Adaptation of a Framework for Dynamic Adaptation of Services

Françoise André, Erwan Daubert, Guillaume Gauvrit

IRISA / INRIA

Campus de Beaulieu, 35042 Rennes cedex, France

{Francoise.Andre, Erwan.Daubert, Guillaume.Gauvrit}@irisa.fr

Abstract—The dynamism and scale of the infrastructure of the Internet of Services bring new needs to build autonomous services. These services have to be able to self-adapt to the variation of the environment. Moreover, these adaptations may span across multiple services and thus have to be coordinated, without breaking their autonomy. To this end we describe in this paper the approach we have chosen for SAFDIS, a framework to make coordinated adaptations of services. In this presentation, a particular emphasis is made on the distribution of the framework and how it helps to coordinate distributed adaptation. Benefits from the self-adaptation of the framework itself are also presented.

Keywords—Dynamic Adaptation; Distributed Services; Distributed Adaptation; Self-Adaptation of Adaptation Framework.

I. INTRODUCTION

The underlying computing infrastructure for the Internet of Services is characterized by its very large scale, heterogeneity and dynamic nature. The system scale is to be measured in terms of number of users, services, computers and geographical wingspan. The heterogeneity comes from its spreading on multiple sites in multiple administrative domains providing very different computers, devices and network connections. Its dynamic nature results from a number of factors such as Internet node volatility (due to computer or network failures, voluntarily connections and disconnections), services evolution (services appearing, disappearing, being modified) and varying demands depending on human being activities.

In a world of services in which more and more personal, business, scientific and industrial activities rely on them, it is essential to guarantee the high availability of services despite failures or changes in the underlying continuously evolving execution environment. Moreover, providing quality of service (QoS) is important considering the number of services related to legal and commercial aspects.

To take into account this dynamism our objective is to design and implement systems that are context aware and able to adapt applications and services at run-time.

The task of making software adaptable is very cumbersome and encompasses different levels:

- At user or business level, processes may need to be reorganized when some services cannot meet their Service Level Agreement (SLA).
- At service composition level, applications may have to change dynamically their configuration in order to take into account new needs from the business level or new constraints from the services and the infrastructure level. At this level, most of the applications are distributed and there is a strong need for *coordinated adaptation*.
- At infrastructure level, state of resources (networks, processors, memory, etc.) have to be taken into account by service execution engines in order to make a clever use of these resources, such as taking into account available resources and energy consumption. At this level there is a strong requirement for *cooperation* with the underlying operating system.

Moreover, the adaptations at these different levels need to be coordinated.

So, our main challenge is to build a generic framework for self-adaptation of services and service based applications. The basic steps of an adaptation framework are Monitoring, Analysis, Planning and Execution, following the MAPE model proposed in [1]. We intend to improve this basic framework by refining each step of the MAPE model, in particular by providing elements that cope with the distribution of the application and the underlying infrastructure. The adaptation system can itself be distributed for the purpose of scalability or to better match the heterogeneity of the environment. Moreover, it can be adaptable, allowing to take into account unforeseen situations.

Our system called SAFDIS for Self-Adaptation For Distributed Services fully exploits the advantages of the framework concept [2]. It gives a frame, paradigms and rules to develop and implement adaptation mechanisms, as well as the liberty and the flexibility for the developer to specialize its system according to its specific needs. Using this framework, the task of developing concrete adaptation systems for some applications, services or infrastructures will be facilitated as many of the different elements that may be composed in adaptation systems are exposed, their

interfaces clearly defined, the relationships between them coherently specified. Our SAFDIS framework is build as a set of services, providing functionalities useful to build an adaptation system. Not all functionalities are necessarily needed for each instantiation of an adaptation system. For instance we provide a *negotiator* service to negotiate the adaptation decisions when SAFDIS is distributed on several nodes; this service is not useful when SAFDIS is build as a unique centralized adaptation system.

The following sections present the advantages resulting from the design of our framework. The next section gives an overview of the SAFDIS framework. Section III presents how the distribution is handled in our framework and its advantages. Then, Section IV presents some advantages of having an adaptation framework that is self-adaptable. Finally, Section V presents some related-works and Section VI concludes this paper.

II. SAFDIS: SELF-ADAPTATION FOR DISTRIBUTED SERVICES

Our framework for self-adaptation of distributed services SAFDIS [3] is divided into the four main phases of the MAPE model. *Monitoring* is the observation function to detect changes that imply adaptation. When a change is detected, the monitoring phase triggers the *analysis* to analyze it and find an adaptation strategy if it is required. Then this strategy is given to the *planning* phase to compute a schedule of actions that will implement the strategy. The last step is the *execution* of the schedule to reconfigure the system (application, services and the environment).

Our framework is able to work at different levels ranging from a single service, a composition of services for one application, to several applications. Each application can be executed on a set of heterogeneous platforms themselves on a distributed and heterogeneous infrastructure (OS and hardware). Therefore in order to adapt a set of applications, it may be necessary to interact with these platforms and some specific (maybe all) execution nodes which represent only a part of the infrastructure. With SAFDIS, it is possible to monitor the different levels according to the implementation of the available probes and adapt them depending on the adaptation actions (Figure 1).

To cope with the distributed environment, our framework can itself be distributed using multiple autonomous and cooperating instances. An instance has to be deployed on each of the service oriented platforms hosting a least an adaptive service using SAFDIS. Our framework is also fully decentralized, meaning that there are no instances with privileges or special purposes. This design avoids single points of failure and makes the framework scalable.

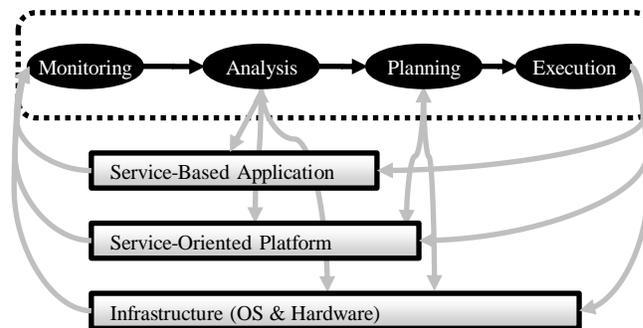


Figure 1. Multi-level Adaptation

In the following subsections, we present each phase of the MAPE model and some of their characteristics in the context of our framework and our current implementation.

A. Monitoring

The monitoring phase is used to provide an informative and dynamic view of the adaptive entity and its environment to the other phases of SAFDIS. Thus, it is the starting point of every adaptation undertaken. It builds one local view per instance of SAFDIS picking relevant information from the service-oriented platform, the adaptive services, the operating system and the hardware. SAFDIS can probe both passively or actively the system to generate events and update the view. The pieces of information that have to be gathered are specified by the other phases of the framework.

B. Analysis

The analysis phase of a MAPE adaptation system has two goals. The first goal is to identify situations needing an adaptation. It listens to updates of the view of the system pictured by the Monitoring phase. Then it analyzes the changes in the system and decides if an adaptation is needed consecutively to this change. The second goal of the analysis phase is to build an adaptation strategy when a need arises. A strategy defines which elements (parameters, functions...) need to be adapted and how.

Within our SAFDIS implementation, the analysis phase takes decisions with multiple temporal scopes. This gives the ability to either react fast or to take proactive decisions for the long term. This implies the ability to analyze the context with a variable depth of reasoning. Our implementation of the SAFDIS analysis phase also distributes and decentralizes its analysis process to spread the computational load and make the analysis process scalable.

C. Planning

The planning phase seeks the set of actions (the plan) needed to adapt the system according to the strategy chosen

by the analysis phase. It also schedules the selected actions to ensure a coherent and efficient execution of the adaptation.

Until now the planning phase has received little attention in the context of adaptation and in many cases the planning algorithms used produce simple total orderings of actions. In these cases, the result is not very efficient since the execution may take more time than necessary as the actions are sequentially executed. Moreover in distributed environments, where actions can be asynchronous, some synchronization actions explicitly have to be added to ensure the predefined sequential order.

The planning topic is a well known subject in AI research works and many algorithms already exist in that field to produce efficient schedules. With our SAFDIS framework, the planning phase is able to reuse these algorithms. The resulting plan of actions can have actions that can be executed in parallel.

D. Execution

Once the planning phase has computed the action plan corresponding to the strategy, the execution phase is called to perform the adaptation actions on the service, the application or the environment. These actions are application, platform, OS or hardware specific. That's why, with SAFDIS, we have introduced two kind of actions. The first kind called *concrete actions* corresponds to the action implementations which are specific to the adapted element. The second kind called *abstract actions* constitutes an abstraction of the concrete actions. This allows the planning phase to work with abstract actions without taking into account their specific implementations and doing so to build generic action plans.

III. DISTRIBUTION

The SAFDIS framework is meant to be distributed in the same way the applications it adapts are distributed. When deployed, it is composed of multiple autonomous instances, each one in charge of the adaptation of the services deployed on its platform. However, these autonomous instances cooperate in order to coordinate the adaptations involving distributed elements.

Moreover it is also fully decentralized: there are no instances with privileges or special purposes, therefore there is no single point of failure. When an instance fails, for example from a hardware failure or power issue, the other instances can continue to operate normally, even though the adaptations related to the failed instance will fail.

The absence of an instance with a central role avoids the bottleneck problems that could arise from this role. Also, there is no need for a server dedicated to the management of the adaptation of the services.

For example, let's consider the adaptive services S_a , S_b and S_c respectively executed in the service-oriented

platforms P_a , P_b and P_c on the execution nodes N_a , N_b and N_c . The service S_a uses the services of S_b and S_c . The three services are using SAFDIS in order to be adaptive, thus there is an instance of SAFDIS on each service-oriented platform: I_a , I_b and I_c . If I_a and I_b are involved in an adaptation on S_a and S_b whereas at the same time the node N_c sees its CPU and memory load decrease, I_c can without having to ask the other SAFDIS instances make the adaptation decision consisting in allocating more memory to S_c thus improving the quality of this service. Both adaptations, the one concerning S_a and S_b and the one concerning S_c , can be executed in parallel.

When deployed, SAFDIS is a set of distributed instances. Each instance is a set of services which fulfill the four main functions of SAFDIS: monitoring, analysis, planning and execution. The cooperation between the instances is done at the level of the services. For example, analysis services cooperate among themselves but none of them interact with the monitoring and planning services that are outside of its SAFDIS instance. This respects the separation of the adaptation process into four phases.

As there is one instance of SAFDIS on each service oriented platform, each monitoring service is in charge of monitoring its execution node, the platform itself, the services deployed on its platform and the SAFDIS instance it is part of. The other services of SAFDIS always send their requests of information to the local monitoring service. This monitoring service then retrieves the information from another monitoring service if it doesn't have it. The connection between the various monitoring services are made on demand, using a service registry.

Instead of trying to picture a global view of every elements contributing to the application, which would consume communication resources and not be scalable, SAFDIS pictures multiple local views. This allows to spread the computation load of the analysis on the execution nodes related to the adaptations. But this means that the instances of the analysis component have to take decisions based on partial knowledge of the system. This knowledge alone is not always enough to decide of adaptation strategies. Thus, the analysis instances use negotiation mechanisms in order for them to cooperate in the decision process.

The analysis services cooperate by negotiating strategies. A strategy is initiated by an analysis instance and then negotiated with the other analysis instances that are (or may be) impacted in the adaptation. Those instances can enhance the proposed strategy. They may in turn involve other analysis instances into the negotiation process.

In the previous example involving three adaptive services and SAFDIS instances, if I_a chooses a strategy and negotiates it with I_b and I_c , the last two instances analyze in parallel the portion of the strategy requiring their involvement.

I_a takes the final decision to apply the negotiated strategy or to abandon it.

Each analysis service uses three sub-services: a decision maker service in charge of the reasoning and decision making process and a pair of services to handle the negotiation: the negotiation manager and the negotiator services. Each negotiator handles one to one negotiations while the negotiation manager divides the negotiations involving more than two peers into multiple negotiations involving two peers and coordinates them. This design was chosen to enforce a separation of concerns and to ease potential upgrades when SAFDIS is deployed. In our implementation of the framework, the negotiation protocol used by the negotiation manager and negotiator services is the Iterated Contract Net Protocol [4]. However, this is transparent for the decision maker service, so other negotiation protocols could be used.

Once a strategy is negotiated, it is sent to the planning service that is in charge to make a plan of actions to implement it. As said in II, due to the planning algorithms used in SAFDIS, some actions can be scheduled to be executed parallel.

So, in the last phase of the adaptation process which is the effective execution of the adaptation, the distributed aspect of the adaptation process is again emphasized as the actions that can be executed in parallel are executed in parallel, which in a distributed context improves the execution time of the adaptation.

Overall, the distribution of the analysis process and the distribution and the parallelism of the execution phase allows our framework to spread the computation load of the adaptation process and to gain time in this process.

IV. SELF-ADAPTATION OF THE ADAPTATION SYSTEM

As our SAFDIS framework is developed as a service-oriented application it can itself be adapted as any other application, using its own mechanisms. In this section we present this self-adaptation property and detail its use for the planning phase.

The current implementation of each MAPE phase of SAFDIS can be dynamically replaced by a new implementation. At deployment a first implementation for each phase is chosen by the expert. If the expert think that there is no chance that the context will necessitate his choice to be called into question, SAFDIS will remain the same a long time. But the expert can predict that his initial choice may not be the best one in every circumstances or if some changes appear in the future. In that case he can add policies to the SAFDIS framework he initially used to adapt the application, in order to adapt the framework itself.

Thanks to the use of a service based adaptation framework design, the need to stop the application and its execution environment when changing the adaptation system is avoided.

The adaptation system itself needs not to be completely stopped as a phase may be changed without affecting the others. The expert only has to have foreseen other implementations for the phase subject to potential changes and the policies to decide the change. Of course the new implementation should respect the services specifications, especially be conform with the interfaces defined in our framework to ensure the communication between the MAPE phases. At run-time, the new implementation will be looked for in the services repository, started and then the previous one will be stopped. Interconnections between phases are automatically done through the interfaces, without the help of the expert.

Portions of a phase are also self-adaptable without having to replace the complete phase. This is the case for instance of the negotiation part of the analysis phase.

To illustrate this self-adaptation property, we detail below the way it has been conceived and developed in our current prototype for the Planning phase.

As adaptation is performed at run-time, the time needed to actually perform the adaptation have to be minimized. Therefore, planning is an important phase of the MAPE model. It chooses the actions necessary to properly apply the adaptation strategy, and schedules the actions to ensure the consistency of the adaptation execution.

A simple planning algorithm as used by most adaptation systems uses a static total ordering between all possible actions and leads to a sequential schedule.

For example, if we consider the three possible actions *stop service*, *update service* and *start service* and an order that imposes that whatever the number of services to change all the *stops* must be done before the *updates* and all *updates* before the *starts*, this adaptation method will maximize the time during which all the services are unavailable, and also consume more time than needed in case some actions may have been processed in parallel.

Moreover, if the adaptation takes place on a distributed and asynchronous environment, explicit synchronization operations should be added to enforce the respect of the schedule between the different parts of the actions that have to be executed on different platforms.

Research works on planning methods such as Artificial Intelligence planning, Motion planning or Control theory, have produced algorithms [5], [6] that overcome these limitations, but without applying them in the context of dynamic adaptation. In SAFDIS, we propose an architecture for the planning phase (called F4Plan for "Framework For Planning adaptation" [7]) that offers the possibility to use, according to the needs, one of these algorithms. Moreover this architecture includes a set of language translators that allow to translate the possible output languages from the analyze phase (languages used to describe the current configuration

of the application and the target configuration) into the different input languages used by the planning algorithms.

The self adaptation of the planning phase consists in choosing the most suitable planning algorithm according to policies defined by the expert of adaptation. These policies are based on some non-functional constraints defined by the system such as the system overload or the duration which may be acceptable in order to apply the strategy but they also take into account the strategies sent by the analysis phase. For example, if the strategy comes from a reasoning engine that is used to do local adaptations to solve local problems, such as the one we use to make reactive decisions based on event-conditions-actions rules, it is not necessary to use a planning algorithm that searches for a parallel schedule. Indeed, there will probably be relatively few actions to schedule and all of them should be executed on the local node. In that case the simplest planning algorithm is convenient, being able to plan the strategy as quickly as possible, thus minimizing the time spent in the planning phase.

At the opposite, if the strategy comes from a reasoning engine based on utility functions such as the we use to make proactive decisions to do wide adaptations impacting the distributed system, it is interesting to use a planning algorithm able to plan a strategy as efficiently as possible. This planning algorithm should take into account several constraints for example the potential asynchronism between actions and the amount of resources that will be used during the execution phase.

So, the modularity and the service based design of our SAFDIS framework allows a great flexibility in the conception of an adaptive system. We do not neglect of course the task of the adaptation expert who has to conceive the adaptation policies.

V. RELATED WORKS

Today research works on autonomic computing aim mainly to build autonomic components but very few works consider building autonomic services or autonomic service-based applications. Among these works most of them as [8], [9] integrate the adaptation process into the components or services. Each element constitutes an autonomous element of the system and it doesn't interact with other elements to coordinate more complex adaptations. So, these solutions are not appropriate to manage wider adaptation spanning over multiple services constituting one or more applications. Meanwhile in [8], the authors add some predefined high-level adaptation components to be able to adapt a set of elements constituting an application. But this possibility is restricted to some specific cases for example to resource management or application deployment.

Other works as in [10], [11] separate the behavior of the components or services from the adaptation process. In [10] the generic framework called Dynaco needs to be specialized and is specific for each application, so several instances of the Dynaco framework are needed to adapt multiple applications.

Among these solutions, very few manage distributed systems and are themselves distributed. Based on Dynaco, [12] proposes some coordination patterns to cope with the distribution and decentralization of the adaptation system. However, to our knowledge these solutions are not able to manage heterogeneous applications.

VI. CONCLUSION

Nowadays, software developments should consider the issue of their adaptation when confronted with the dynamism of execution environments. However current solutions for adaptation are most often ad hoc and in consequence are not satisfying as long term solutions.

With our framework, which targets service-based applications, we propose to externalize the adaptation process into a distinct and distributed application. This new application is able to interact with various heterogeneous applications, services and execution platforms to adapt them. Moreover, as a distinct application it is able to adapt itself. In this paper, we have described some characteristics and advantages of our SAFDIS framework to ease the design of adaptation systems for service-based distributed applications. Some relevant parts of our implementation have also been presented.

Our framework provides a set of interfaces useful to build an adaptation system including some optional functionalities, such as a negotiator service which is used to negotiate the adaptation decisions when SAFDIS is distributed on several nodes. It is the role of an expert designer who knows his application and the execution environment to specialize our framework and to choose whether to use those optional functionalities. Moreover, our implementation is built as a Service-Based Application in order to take advantage of the service-oriented approach. For example, the dynamic binding between services eases the replacement of services when updating some part of the adaptation system. We also integrate some self-adaptation capabilities in our adaptation system and use them to select the planning algorithm.

The SAFDIS framework has been experimented to adapt test applications such as video streaming and multi-support video conferences applications. The planning phase has been used for the adaptation of an home-automation application [7], showing significant improvement compared to the initial version of the adaptation system. We are currently working on the design of the adaptation system for a large and very dynamic firemen assistance application.

In order to improve our implementation, we plan to study the use of already defined planning algorithms which are able to distribute the planning process ([13], [14], [15]) and to integrate them. This will help distribute the computation load in the same way it helps the analysis process. We also plan to work on conflicts that may appear in simultaneous adaptation processes. This kind of conflicts may appear because since a distributed and decentralized adaptation system is used, many adaptation processes may be launched and these processes may have to adapt the same element. In that case one of those adaptations may fail or may be inefficient. A third point we plan to study is the use of a knowledge base to share data between adaptation phases and to build a history about the system. This history may be used to improve the quality of the analysis phase by providing feedback on previous adaptations and to ease the resolution of conflicts by providing some information about the state of the running adaptations.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement 215483 (S-Cube).

REFERENCES

- [1] S. R. White, J. E. Hanson, I. Whalley, D. M. Chess, and J. O. Kephart, "An architectural approach to autonomic computing," *Autonomic Computing, International Conference on*, pp. 2–9, 2004.
- [2] D. Riehle and T. Gross, "Role model based framework design and integration," in *In OOPSLA '98: Proceedings of the 1998 Conference on Object-Oriented Programming Systems, Languages, and Applications*. ACM Press, 1998, pp. 117–133.
- [3] G. Gauvrit, E. Daubert, and F. André, "SAFDIS: A Framework to Bring Self-Adaptability to Service-Based Distributed Applications," in *36th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA)*, Lille France, 09 2010, pp. 211–218.
- [4] "Fipa interaction protocol specifications," 2010. [Online]. Available: <http://www.fipa.org/repository/ips.php3>
- [5] A. L. Blum and M. L. Furst, "Fast planning through planning graph analysis," *Artificial Intelligence*, vol. 90, pp. 1636–1642, 1995.
- [6] Y. Chen, C. wei Hsu, and B. W. Wah, "Sgplan: Subgoal partitioning and resolution in planning," in *In Edelkamp*, 2004, pp. 30–32.
- [7] F. André, Daubert Erwan, Nain Grégory, M. Brice, and B. Olivier, "F4Plan: An Approach to build Efficient Adaptation Plans," in *7th International ICST Conference on Mobile and Ubiquitous Systems (MobiQuitous)*, Sydney, Australia, December 2010.
- [8] D. M. Chess, A. Segal, and I. Whalley, "Unity: Experiences with a prototype autonomic computing system," in *ICAC '04: Proceedings of the First International Conference on Autonomic Computing*. IEEE Computer Society, 2004, pp. 140–147.
- [9] J. Sudeikat, L. Braubach, A. Pokahr, W. Renz, and W. Lamersdorf, "Systematically engineering self-organizing systems: The sodekovs approach," in *Proceedings des Workshops über Selbstorganisierende, adaptive, kontextsensitive verteilte Systeme (KIVS 2009)*. Electronic Communications of the EASST, 3 2009, p. 12.
- [10] J. Buisson, F. André, and J. Pazat, "Supporting adaptable applications in grid resource management systems," in *Proceedings of 8th IEEE/ACM International Conference on Grid Computing*, Austin, USA, 2007, pp. 58–65.
- [11] P. C. David and T. Ledoux, "An Aspect-Oriented approach for developing Self-Adaptive fractal components," in *Software Composition*, ser. LNCS, vol. 4089, 2006, pp. 82–97.
- [12] M. Zouari, M.-T. Segarra, and F. André, "A framework for distributed management of dynamic self-adaptation in heterogeneous environments," in *The 10th IEEE International Conference on Computer and Information Technology*, Bradford Royaume-Uni, 06 2010, pp. 265–272. [Online]. Available: <http://hal.inria.fr/inria-00471892/en/>
- [13] F. Gechter, V. Chevrier, and F. Charpillet, "A reactive agent-based problem-solving model: Application to localization and tracking," *ACM Trans. Auton. Adapt. Syst.*, vol. 1, no. 2, pp. 189–222, 2006.
- [14] M. P. Georgeff, "Distributed artificial intelligence," A. H. Bond and L. Gasser, Eds. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988, ch. Communication and interaction in multi-agent planning, pp. 200–204.
- [15] P. Buzing, A. T. Mors, J. Valk, and C. Witteveen, "Coordinating self-interested planning agents," *Autonomous Agents and Multi-Agent Systems*, vol. 12, no. 2, pp. 199–218, 2006.

Minimizing Human Interaction Time in Workflows

Christian Hiesinger, Daniel Fischer, Stefan Föll, Klaus Herrmann, Kurt Rothermel
University of Stuttgart

Institute of Parallel and Distributed Systems (IPVS)
Universitätstr. 38, D-70569 Stuttgart

{chrisitan.hiesinger,daniel.fischer,stefan.foell,klaus.herrmann,kurt.rothermel}@ipvs.uni-stuttgart.de

Abstract—Many business scenarios require humans to interact with workflows. To support humans as unobtrusively as possible in the execution of their activities, it is important to keep the interaction time experienced by humans as low as possible. The time required for such interactions is influenced by two factors: First, by the runtime of the services that are used by a workflow during an interaction. Second, by the time required to transfer data between workflow servers and services that may be distributed in a global network. We propose an algorithm that computes a suitable distribution of a workflow in such a network. The goal of our algorithm is to minimize the time required for interactions between a human and a workflow. Current approaches in the domain of workflow optimization pay little attention towards optimizing a workflow to increase the usability for humans. We show the feasibility of our approach by comparing our algorithm with two non-distributed approaches and a distributed approach which is based on a greedy algorithm and show that our algorithm outperforms these approaches.

Index Terms—Workflow distribution, human interaction, pervasive workflows.

I. INTRODUCTION

By using workflows, organizations are able to automate and optimize their business processes [1]. In many business scenarios, activities have to be executed by humans. Therefore, it is important to integrate humans into workflows.

Such integration can adhere to different patterns. In the simplest of cases, a human is notified by the workflow system about currently available activities and provides some sort of feedback when he has completed them. However, more complex interaction patterns may require a human to query the workflow system for information throughout the execution of the activities. The time needed to provide the desired information is experienced by the human as waiting time. An important design principle from the area of Pervasive Computing is to support humans as unobtrusively as possible [2], [3]. Therefore, applying workflows in this area requires that such waiting times are minimized.

The time is dependent on two factors: First, on the runtime of services that need to be executed in order to provide a human with the desired information. Second, on the time required to transfer data between workflow servers and service hosts participating in the execution of a workflow. New technologies like Cloud Computing support organizations in focusing on their core business [4] and lead to the necessity of using remote

service providers within workflows. However, communicating data to remote networks may create extensive waiting times due to limited bandwidth and significant propagation delay.

In this paper, we propose an algorithm for distributing a workflow over a set of workflow servers such that the interaction time experienced by humans is minimized. Existing approaches for workflow optimization do not take this factor into account [5], [6], [7]. Our algorithm is based on a two-phase list-scheduling approach. In phase 1, an initial distribution is computed that is based on estimated values for activity execution and data transfer times. In phase 2, the initial solution is refined based on a hill-climbing algorithm. We show that our algorithm improves the interaction time between humans and workflows by up to 80% compared to an approach in which the complete workflow is run on a single machine. Furthermore, we report an improvement of up to 10% compared to an existing greedy approach. We also show that our algorithm scales better with an increasing number of tasks compared to this approach.

The remainder of this paper is organized as follows. In Section II, we introduce our system model and define the problem of workflow placement in a formal way. Thereafter, in Section III, we discuss related work in the area of workflow placement. In Section IV, we describe our placement algorithm that we evaluate in Section V. Finally, we conclude the paper and give some outlook on future work in Section VI.

II. SYSTEM MODEL AND PROBLEM DESCRIPTION

In this section, we describe our system model in a formal way. Our goal is to distribute a workflow among local networks (*domains*) administered by various service providers which are connected to each other via a global network. We split our system model into a network model and a workflow model. Thereafter, we formalize our goal of minimizing the time to execute so called *human interaction patterns* as an optimization problem.

A. Network Model

We assume a set of domains D , each representing a local network consisting of a set of hosts. A domain $d \in D$ provides a set of *service types* S_d . Services of the same type may be available (replicated) in different domains and $S = \bigcup_{d \in D} S_d$. Note that we model the functionality a human h provides as a special service $s_h \in S$. We assume a workflow server and a *domain controller* in each domain. The

This research has been supported by FP7 EU-FET project ALLOW (contract number 213339).

domain controller serves as an information service. It provides the link properties of all relevant communication links and a service discovery mechanism. This can be achieved by means of an overlay network between all domain controllers in the network. Services, workflow server and domain controller may be replicated inside a domain to allow for provisioning of quality of service guarantees, but for simplicity we treat each of them as being unique within the respective domain.

We assume that each domain is able to communicate to any other domain via the Internet. We denote the bandwidth and propagation delay between two arbitrary domains $d_1, d_2 \in D$ as $\beta(d_1, d_2)$ and $\delta(d_1, d_2)$, respectively. The bandwidth and propagation-delay between two hosts in the same domain is assumed to be constant and denoted as $\beta(d, d)$ and $\delta(d, d)$, respectively. We assume that $\forall d \in D, \forall d_i, d_j \in D, d_i \neq d_j : \beta(d, d) \gg \beta(d_i, d_j) \wedge \delta(d, d) \ll \delta(d_i, d_j)$, i.e. inter-domain delay dominates intra-domain-delay which reflects typical communication properties found in interconnected LANs.

B. Workflow model

A workflow is a directed acyclic graph $W = (A, s, C, \rho, \theta_A, \theta_D)$. A denotes the set of activities in the workflow. The functionality of an activity is defined by means of the function $s : A \rightarrow S$ which maps an activity $a \in A$ to a required service type $s \in S$.

The control flow is specified by means of the set $C \subset A \times A$ and defines the logical order of activities. We refer to activities that model conditional or parallel behaviour as *structural activities*. A conditional and parallel split is modeled as an activity with more than one outgoing control flow link. The set of outgoing control flow links of an activity $a \in A$ is denoted as C_a . For a given control flow link $c = (a_i, a_j)$, ρ_c is the probability that a_j will be executed if a_i has been executed. This value can be derived from execution traces of the workflow. For a conditional split, the workflow is executed following only a single alternative, i.e. the conditions $|C_a| > 1$ and $\sum_{c \in C_a} \rho(c) = 1.0$ hold. For a parallel split, all outgoing branches are executed in parallel, i.e. the conditions $|C_a| > 1$ and $\forall c \in C_a : \rho_c = 1$ hold. The latter also holds for all other links originating from a non-structural activity.

The average amount of data that needs to be transferred from a workflow server executing an activity $a \in A$ to a service required by a is denoted as $\theta_S(a)$. We assume that the values of $\theta_S(a)$ cover the amount of data required for the input parameters as well as for the result of the respective service call. Similarly, $\theta_A(a_1, a_2)$ specifies the average amount of data that has to be exchanged between two activities $a_1, a_2 \in A$. We assume that the functions θ_A and θ_S are defined either by means of estimations by a workflow designer or by learning them from past executions of the workflow. In the following, we refer to communication relationships between activities as well as between an activity and a service as *data links* and subsume them in the sets $L_{AA} \subset A \times A$ and $L_{AS} \subset A \times S$, respectively.

A *Human Interaction Pattern* (HIP) is a connected subgraph of a workflow. It starts with a single entry activity which

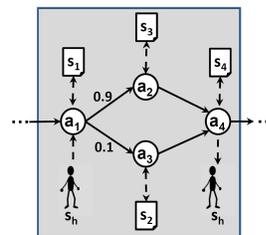


Fig. 1. Human Interaction Pattern

expects input data from a human and ends with a single exit activity which generates output data for the same human. This is a natural assumption as we focus on interaction patterns that resemble queries. An example for a HIP is given in Figure 1. Arrows show control flow links while circles and rectangles represent activities and services, respectively. The single entry activity is a_1 (a conditional split). The single exit activity is a_4 . Note that there may be several HIPs in a single workflow.

C. Problem description

Our goal is to find a mapping function $\mu : A \rightarrow D$ of activities to domains that minimizes the average required communication time for all HIPs in a workflow. We focus only on activities that are part of a HIP. All other activities may be distributed according to other optimization goals (e.g. network load). Each execution of the workflow takes only a single *route* through the workflow. A route is a connected subgraph of a workflow that contains all activities visited in one execution. For example, a_1, a_2, a_4 as well as a_1, a_3, a_4 are both valid routes in the HIP shown in Figure 1.

Because we do not know this route in advance, we cannot optimize our mappings for it. Therefore, we solve the most general case and aim for minimizing the expected execution time of a HIP. Let R be the set of all routes of a HIP. The probability for the execution of $r \in R$ can be calculated as $q_r = \prod_{c \in r} \rho_c$. Furthermore, let φ_r^μ be a function that defines the execution time for r under the mapping μ . Then, our goal is to find a mapping μ such that the following sum is minimized:

$$\sum_{r \in R} q_r \cdot \varphi_r^\mu \quad (1)$$

In the following, we describe how φ_r^μ is calculated. In a parallel split, only the branch which results in the largest execution time is relevant for the overall execution time of the flow. We refer to the subgraph of r that contains only this longest branch for every parallel split as the *critical path* of a route. The time to execute a route of a HIP is influenced by three factors: The time κ_A required to transfer data between the activities, the time κ_S required to transfer data between activities and their mapped services, and by the time κ_X required to execute the corresponding services. Thus, we have $\varphi_r^\mu = \kappa_A + \kappa_S + \kappa_X$.

For the computation of κ_A , we have to distinguish two cases. First, two activities that exchange data may be mapped to the same domain and, thus, to the same workflow server according to our system model. In this case, κ_A is negligible because no data has to be sent over the network. Second, two

activities may be mapped to different domains. In this case, the time required equals the sum of the propagation delay of the communication link between the respective domains and the time required for transmitting the data on the respective data flow link. Let $L_{crit} \subseteq L_{AA} \cup L_{AS}$ be the set of data links on the critical path of a route r , then

$$\kappa_A = \sum_{\forall (a_i, a_j) \in L_{crit}} \delta(\mu(a_i), \mu(a_j)) + \frac{\theta_A(a_i, a_j)}{\beta(\mu(a_i), \mu(a_j))}. \quad (2)$$

For the computation of κ_S , we proceed analogously. Given an activity a placed in domain d , let $\xi(a)$ be a function that returns the domain, among all domains that provide an instance of service type $s(a)$, which can be accessed with lowest interaction time (ideally, $\mu(a) = \xi(a)$). Then,

$$\kappa_S = \sum_{\forall (a, s) \in L_{crit}} \delta(\mu(a), \xi(a)) + \frac{\theta_S(a)}{\beta(\mu(a), \xi(a))}. \quad (3)$$

We assume that service providers guarantee a certain execution time as part of a SLA. For the computation of κ_X , we accumulate the expected runtime required for the services mapped to the activities on the critical path.

Our problem is a generalization of the problem of task allocation in heterogeneous distributed systems (TAHDS) which is known to be NP-hard [8].¹ Therefore, we propose to use a heuristic algorithm to solve the problem because an exhaustive search of an optimal placement for example by means of backtracking is not feasible.

III. RELATED WORK

Bauer and Dadam [6] propose an algorithm for assigning workflow activities to servers in order to reduce network load. They introduce a cost model based on estimated execution data and employ a greedy algorithm. First, each activity is greedily placed on a workflow server that minimizes the cost for its execution. Then, a hill-climbing algorithm is used to eliminate data transfers between neighbouring activities which have been placed on different servers. In this approach activities are initially placed without taking their data links into account. Thus, it is unlikely that suitable sets of service providers are found in our scenario. We will show that our heuristic performs better than a version of this algorithm adapted to our problem. We refer to this adapted version as *Greedy* approach.

Son et al. [5] propose an algorithm for minimizing communication cost based on multi-level graph partitioning. A workflow is divided into several fragments. However, their solution assumes homogeneous communication links which is not valid in the Internet.

In parallel computing, tasks have to be assigned to CPUs in order to optimize their execution. Many solutions assume that activities are not depending on each other, which leads to a Bin-Packing problem. Obviously, this assumption does not hold for our problem. More sophisticated approaches apply list scheduling algorithms [8], [9]. We adopted the basic idea

of these algorithms while dropping their basic assumption of homogeneous communication links.

In the area of grid computing, list scheduling algorithms are employed under the assumption of heterogeneous network links among loosely coupled computing systems [7], [10]. However, these algorithms assume variable task execution times to have a major influence on the overall execution of the task graph. In our scenario, we assume that tasks are services and that quality of service guarantees specify the time required for their execution. Hence, in our case the overall execution time mainly depends on the communication links between workflow servers, rendering respective algorithms like e.g. HEFT inappropriate.

IV. HEURISTIC PLACEMENT ALGORITHM

We propose a 2-phase algorithm based on a list-scheduling approach in order to find a mapping μ that minimizes the runtime of HIPs. Since HIPs are independent of each other, we map each HIP separately.

As soon as activity a is mapped to domain d , activities with a communication dependency to a have to communicate with d . Thus, they should not be assigned to the best domains in a greedy fashion. Instead, we have to take this dependency into account and map each activity depending on its influence on the runtime of a HIP: The more influence it has on the runtime, the earlier it is mapped.

According to our system model, the bandwidth and propagation delay between domains vary and the time required for communication depends on the network link used. Additionally, there may be services which are available in many domains while other services are only available in few domains. Therefore, it is not known which data link is the most expensive (in terms of communication time) until all activities have been mapped. Therefore, we use a heuristic to estimate the cost of each data link before the actual mapping. Since HIPs are independent of each other, we run the algorithm for each HIP in a workflow as soon as it is known from which domain the human accesses the workflow.

In a first phase, we assign weights to the data links to reflect their estimated costs. Then, we sort the links in descending order of their weights to ensure expensive activities are mapped to domains first. To derive an initial mapping, we iterate over the sorted list and map the activities to domains such that their execution time is minimized. The final mapping is created by optimizing the initial mapping through hill-climbing. The overall algorithm (called *Link Weight Activity Assignment* (LWAA)) is depicted in Listing 1.

A. Weighting and Ordering

The weight of a data link l for the initial mapping is calculated by virtually placing l on each of the possible network links between any pair of domains and by calculating the average time consumed over all these virtual mappings.

Let $l_{AA} = (a_i, a_j)$ be a data link between two activities and let $l_{AS} = (a, s(a))$ be a data link between an activity and its required service. We distinguish between the average time

¹see Appendix for proof of problem complexity

Listing 1 LWAA Algorithm

```

1: // Let  $\hat{\mu}$  be the associative array that represents  $\mu$ 
2: // At the beginning  $\forall a : \hat{\mu}(a) = \perp$  holds
3:  $L_{DF} = \text{weightDataLinks}(L_{AA} \cup L_{AS})$ 
4: /* Initial mapping */
5: while  $L_{DF} \neq \{\}$  do
6:    $l = \text{Link with highest weight in } L_{DF}$ 
7:   if  $l \in L_{AA}$  then
8:      $\text{handleActivityToActivityDataLink}(l, L_{DF}, \hat{\mu})$ 
9:   else if  $l \in L_{AS}$  then
10:     $\text{handleActivityToServiceDataLink}(l, L_{DF}, \hat{\mu})$ 
11:   end if
12:    $L_{DF} = L_{DF} \setminus \{l\}$ 
13: end while
14: /* Optimize mapping */
15:  $\text{hillClimbing}()$ 
    
```

$\text{weight}_W(l_{AA})$ required to communicate between workflow servers and the average time $\text{weight}_S(l_{AS})$ required to access a service from a workflow server that controls the corresponding activity. To compute $\text{weight}_W(l_{AA})$, we consider every possible mapping of two activities a_i and a_j :

$$\text{weight}_W(l_{AA}) = \frac{1}{|D|^2} \sum_{\forall d_k, d_l \in D: k \neq l} \frac{\theta_A(a_i, a_j)}{\beta(d_k, d_l)} + \delta(d_k, d_l) \quad (4)$$

Note that if both activities are mapped to the same domain, no data needs to be transferred.

Similarly, we compute an estimate of the delay created by service data links. In this case, we consider all possible mappings and calculate the average transmission time:

$$\text{weight}_S(l_{AS}) = \frac{1}{|D|} \sum_{\forall d \in D} \frac{\theta_S(a)}{\beta(d, \xi(a))} + \delta(d, \xi(a)) \quad (5)$$

The resulting list of data links is sorted in descending order.

B. Initial mapping

For the initial mapping of each activity to a domain, the algorithm proceeds through the list of data links, in descending order of their weights and maps each activity that has not already been processed. Links connecting two activities (L_{AA}) and links connecting an activity to a service (L_{AS}) are handled differently (cf. Listing 1 lines 8 and 10).

Listing 2 shows the handler procedure for links $(a, s) \in L_{AS}$. This handler finds the domain d that exhibits the least cost for calling service s residing in d when placing activity a in d . Listing 3 shows how to handle a data link $(a, a) \in L_{AA}$. We aim at placing both activities in the same domain such that no data has to be transferred over the network. However, we also do not want to reduce the degree of freedom for the placement more than required. We have to distinguish three

Listing 2 Handle activity to service data link

```

procedure  $\text{handleActivityToServiceDataLink}(l, L_{DF}, \hat{\mu})$ 
    
```

```

1:  $(a, s) := l$  //get corresponding service and activity
2:  $\hat{\mu}(a) := \text{MinArg}d \in D: s \in S_d \delta(d, d) + \frac{\theta_S(a)}{\beta(d, d)}$ 
    
```

Listing 3 Handle activity to activity data link

```

procedure  $\text{handleActivityToActivityDataLink}(l, L_{DF}, \hat{\mu})$ 
    
```

```

1:  $(a_i, a_j) := l$  //get activities the data link connects
2: if  $(\hat{\mu}(a_i) = \perp) \wedge (\hat{\mu}(a_j) = \perp)$  // Both activities unmapped then
3:   if  $\exists d \in D : s(a_i) \in S_d \wedge s(a_j) \in S_d$  then
4:      $a' := \text{Merge}(a_i, a_j)$ 
5:     // Sorted insert of new service data link
6:      $L_{DF} := L_{DF} \cup \{(a', s(a'))\}$ 
7:     // remove service links of  $a_i$  and  $a_j$ 
8:      $L_{DF} := L_{DF} \setminus \{(a_i, s(a_i)), (a_j, s(a_j))\}$ 
9:   end if
10: else if  $(\hat{\mu}(a_i) \neq \perp) \wedge (\hat{\mu}(a_j) = \perp)$  //First activity mapped then
11:   if  $s(a_j) \in S_{\hat{\mu}(a_i)}$  then
12:      $\hat{\mu}(a_j) := \hat{\mu}(a_i)$ 
13:   end if
14: else if  $\hat{\mu}(a_i) = \perp \wedge (\hat{\mu}(a_j) \neq \perp)$  //Second act. mapped then
15:   if  $s(a_i) \in S_{\hat{\mu}(a_j)}$  then
16:      $\hat{\mu}(a_i) := \hat{\mu}(a_j)$ 
17:   end if
18: end if
19: // If both activities are mapped nothing has to be done.
    
```

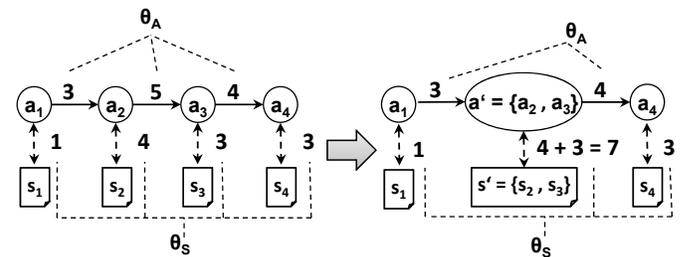


Fig. 2. Merging of unassigned activities

different cases: 1) None of the activities is mapped 2) Only one of the activities is mapped 3) Both activities are mapped.

The first case is handled in lines 2–9 of Listing 3: we check if there exists a domain hosting both service types required by the activities. If such a domain exists, we merge both activities.

The procedure of merging is depicted in Figure 2. The result of merging two activities $a_i, a_j \in A$ is a new activity a' with a corresponding data link to a virtual service $s(a') = s'$ which serves as a container for both $s(a_i)$ and $s(a_j)$. Each operation performed on s' has to be performed for all services in s' . This is illustrated in Figure 2. a_2 and a_3 are merged into a new activity a' with a data link to service type $s' = \{s_2, s_3\}$. The weight of the newly created data link is the sum of the weights of the original links. All other links remain unchanged. Note, that we do not map the merged activities to a domain right away as it may be merged with further activities.

We only merge if there exists a domain hosting the services required by *both* activities because the service access of a_2 and a_3 needs to be restricted to their own domain in order to save communication time. Using the workflow in Figure 2 (left), we explain the rationale behind this idea. Assume that $D = \{d_1, d_2\}$ with $S_{d_1} = \{s_1, s_2\}$ and $S_{d_2} = \{s_3, s_4\}$ and none of the activities is currently assigned to any domain. According to the ranking of data links, the link between (a_2, a_3) has to be processed first. If created a merger $a' = \{a_2, a_3\}$, we would have to map a' either to d_1 or d_2 . Thus, either (a_2, s_2) or (a_3, s_3) would be mapped to an inter-domain link

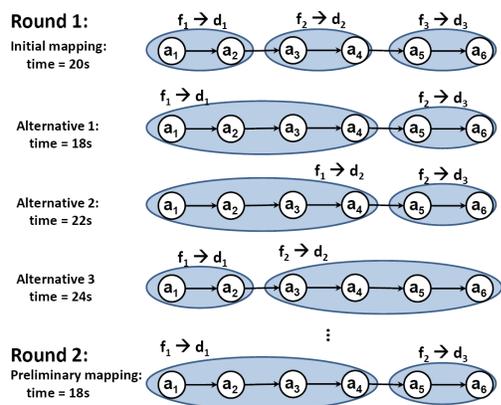


Fig. 3. Hill-climbing

because there exists no domain hosting s_2 and s_3 . The data transferred via the link (a_2, a_3) is either the output data of s_2 or the input data for s_3 . Consequently, we would omit the time required to transfer the data between both activities. However, we would have to transmit the same amount of data via a communication link in the global network which would require the same amount of time that has been saved. Furthermore, through merging in this case we would limit the degree of freedom as, afterwards, we could not map a_2 and a_3 separately.

It may happen that only one of both activities is already mapped to a domain. This is covered in lines 10 to 18 of Listing 3. Analogously to the previous case, we map the unmapped activity to the domain of the already mapped activity only if this domain hosts the required service. Finally, it is also possible that both activities are already mapped to a domain. In this case, we do nothing since the currently handled data link must have a lower priority than the data links that led to a mapping of the respective activities to domains.

C. Optimized mapping

After the initial mapping is completed, we adjust it to the actual bandwidth/propagation delay in the network using a hill climbing algorithm in order to further reduce the time consumed by transferring data via the global network. The principle of the algorithm is depicted in Figure 3.

First, we extract the clusters of the initial mapping. A cluster $A_F \subset A$ is the largest set of activities that form a connected graph with $\forall a_i, a_j \in A_F : \mu(a_i) = \mu(a_j)$. In Figure 3, there exist three clusters in the initial mapping, namely f_1 , f_2 and f_3 . We calculate the time required for the HIP if all activities of a cluster are mapped first to the domain of its preceding and then to the domain of its succeeding cluster. The possible alternatives and the time required for each alternative are depicted in rows 2 to 4 of Figure 3. The best alternative is selected as new preliminary mapping. This procedure is repeated until no mapping which requires less time can be found.

We only remap complete clusters because it is unlikely that remapping single activities results in a performance gain. If

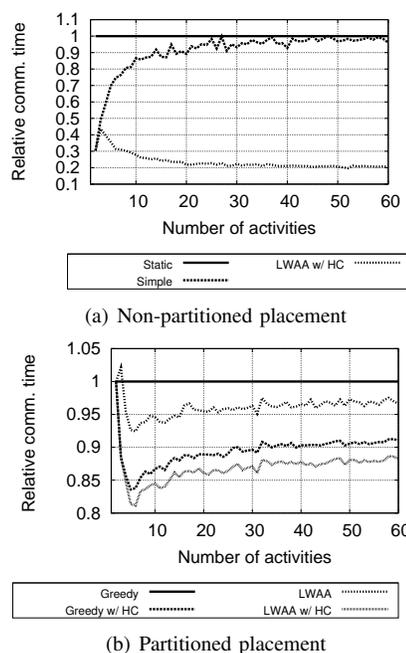


Fig. 4. Comparison with other placement approaches

this would be the case the initial mapping would have come to a different conclusion.

V. EVALUATION

In this section, we describe our evaluation setup and results. We generate networks as well as workflows according to our models discussed in Section II. As the algorithm is executed for each HIP separately, we restrict the generation process to a workflow consisting of a single HIP. The number of activities between the human activities varies between 0 and 60.

We simulate 50 different domains. The bandwidth for communication within each domain is set to 1 GBit/s assuming a Gigabit Ethernet. For the communication between humans and workflow servers, we assume a 54 MBit/s WLAN connection. The bandwidth of communication links between domains is set to be between 2 MBit/s (E1) and 34 MBit/s (E3) to reflect the SLAs between service providers.

We assumed a uniform delay to simplify our simulation. Since the delay between domains is only influencing the ordering of the weighted data links, this does not change the qualitative results.

We have 200 service replicas drawn from 20 services according to a Zipf distribution. The services are randomly assigned to the 50 domains. We use a Zipf distribution because there may be few very popular services available in many domains, while there are many more specialized services which are only provided in a few domains.

For the generation of workflows we use a grammar that is able to generate sequences, conditional and parallel structures. The rules of the grammar are chosen randomly until the desired number of activities is reached. The values for Θ_S are generated randomly according to a uniform distribution with a maximum of 100 MByte to allow for a wide variety of data flow links. The values for Θ_A are implicitly defined by Θ_S to

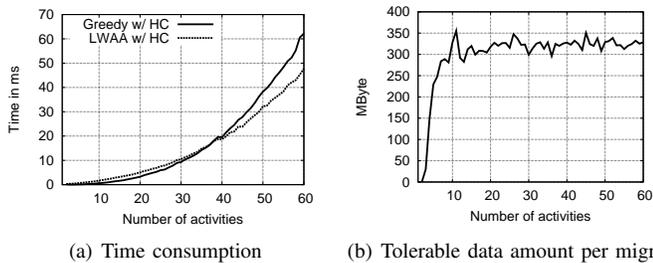


Fig. 5. Performance analysis

guarantee a consistent data flow meaning that all data received by an activity is sent via its outgoing data flow links to other activities. The assignment of activities to (human) services is also chosen uniform randomly.

We compare our algorithm with four other approaches:

- *Static* maps all activities of a HIP to a random domain.
- *Simple* maps all activities of a HIP to the human's current domain.
- *Greedy* proceeds through all activities, searches for the domain reachable with the highest bandwidth hosting the next required service and maps the activity to this domain.
- *Greedy w/ HC* enhance *Greedy* with a subsequent hill-climbing (cf. Figure 3). This is an adapted version of the algorithm proposed by Bauer et al. [6].

In Figure 4(a), comparison of our distribution algorithm (*LWAA*) with the *non-partitioned approaches* (*Static* and *Simple*) is shown. We compared the relative gain of using our algorithm. The reference (at 1.0) is the *Static*. Figure 4(a) shows that, for an increasing number of activities per HIP, our algorithm quickly converges to around 20% of the time required for *Static*. This is because in the non-partitioned approaches a lot of expensive service calls have to use low quality network links. Thus, they consume a lot of time for communication as only the services that are located in the same domain can be accessed in a performant way.

Figure 4(b) depicts the effectiveness of our algorithm compared to the greedy approaches. The initial placement computed by our algorithm is between 12% and 16% better than the placement computed by the *Greedy* approach. This is due to the fact that our algorithm takes the data flow between activities into account and, thus, computes suitable clusters which is not done in the *Greedy* approach. The *Greedy w/ HC* approach performs better than the *LWAA* algorithm without subsequent hill-climbing. This is due to the fact that clusters are assigned to domains without taking the communication links of the individual domain into account. The results show that our algorithm is better compared to the *Greedy w/ HC* approach by 8% to 10% due to the better initial placement.

We compare our algorithm to the greedy approach in terms of the required computation time in Figure 5(a). Both algorithms show very similar execution times at first, until the effort for the subsequent hill-climbing starts to dominate at around 35 activities. This is because of the fact that *LWAA w/ HC* builds activity clusters, reducing the number of clusters

left for the hill-climbing compared to *Greedy w/ HC*.

If the source activity of a data link is mapped to another domain than its target activity, data has to be transferred between workflow servers during the execution of a workflow. This process is called *migration*. The amount of data that has to be transferred in a migration may differ, for example, depending on whether the workflow management system has to transfer additional logging data for compensations. This is not accounted for in our simulation and would impact the performance of our algorithm negatively. Therefore, we measured the amount of data that can be sent additionally for each migration before the *Static* or *Simple* approach outperform our distribution approach. Figure 5(b) shows that this amount can be about three times the maximum amount of data that occurs in the data flow, indicating that considerable migration overhead can be tolerated by our algorithm.

VI. CONCLUSION

We proposed an algorithm that minimizes human interaction time in workflow systems based on a list-scheduling approach for mapping activities to network domains. We compared our algorithm *LWAA w/ HC* to non-partitioned and greedy approaches and showed that it improves interaction time by up to 80%. Hence, *LWAA w/ HC* reduces the interaction time of humans with workflows significantly and, thus, increases the processing throughput considerably. This can result in competitive advantages in a business environment. Additionally, it helps opening areas like pervasive computing for workflow technologies since it renders workflow technology less obtrusive.

In our future work, we will investigate how workflow distribution can help minimizing the energy consumption of mobile devices used for interacting with the workflow. In this case, executing a partial workflow on such a device avoids the energy-intensive transfer of data to the infrastructure.

REFERENCES

- [1] F. Leymann and D. Roller, *Production Workflow: Concepts and Techniques*. Prentice Hall International, 1999.
- [2] M. Weiser, "The computer for the 21st century," in *Scientific American* 265(3): 94-104, 1991.
- [3] S. Schuhmann, K. Herrmann, and K. Rothermel, "A Framework for Adapting the Distribution of Automatic Application Configuration," in *Proc. of the 2008 ACM Int. Conference on Pervasive Services (ICPS 2008), Sorrento, Italy, July 6-10, 2008*. ACM, Juli 2008, Konferenz-Beitrag, pp. 163-172.
- [4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A Berkeley view of cloud computing," University of California at Berkeley, Tech. Rep., February 2009.
- [5] J. H. Son, S. K. Oh, K. H. Choi, Y. J. Lee, and M. H. Kim, "GM-WTA: an efficient workflow task allocation method in a distributed execution environment," *J. Syst. Softw.*, vol. 67, no. 3, pp. 165-179, 2003.
- [6] T. Bauer and P. Dadam, "Efficient distributed workflow management based on variable server assignments," *Lecture Notes in Computer Science*, vol. 1789/2000, pp. 94-109, 2000.
- [7] S. H. H. Topcuoglu and W. M. You, "Performance-effective and low-complexity task scheduling for heterogeneous computing," *IEEE Transactions on Parallel Distributed Systems*, vol. 13, pp. 260-274, 2002.
- [8] Y. Kopidakis, "On the task assignment problem: two new efficient heuristic algorithms," *Journal of Parallel and Distributed Computing*, vol. 42, no. 9, pp. 21-29, 1997.

- [9] A. Billionnet, M. C. Costa, and A. Sutter, "An efficient algorithm for a task allocation problem," *J. ACM*, vol. 39, no. 3, pp. 502–518, 1992.
- [10] A. Radulescu and A. J. C. Van Gemund, "Fast and effective task scheduling in heterogeneous systems," in *Proc. of the 9th Heterogeneous Computing Workshop*. Washington, DC, USA: IEEE Computer Society, 2000, pp. 229–238.

APPENDIX PROBLEM COMPLEXITY

Our problem is a generalization of the problem of task allocation in heterogeneous distributed systems (TAHDS) which is known to be NP-hard [8]. In TAHDS, we have a set of processors P and a set of tasks T . Two arbitrary tasks $i, j \in T$ have communication costs c_{ij} , and e_{ip} represents the cost of executing task i on processor p . The problem is to find a mapping of tasks to processors such that the sum of communication and execution costs is minimized. Note that communication costs between two tasks only occur if they are placed on different processors.

In the following, we reduce TAHDS to our problem in order to show that our problem is NP-hard as well. We map T to A and P to D , i.e. each task corresponds to an activity and each processor to a domain. We define a unique service for each activity and replicate it on every domain. We set the propagation delay between and within domains to zero. Furthermore, we set the bandwidth within each domain to ∞ , i.e. hosts within a domain can communicate instantly. The bandwidth between domains is set to a constant β_{const} . The time to execute service replica $s = s(a)$ running on domain d is set to e_{ip} where a is the activity corresponding to task i and d the domain corresponding to processor p . Θ_A is chosen such that $\Theta_A(a_i, a_j)/\beta_{const} = c_{ij}$ where tasks i and j correspond to activities a_i and a_j , respectively. Obviously, an algorithm that is capable to solve our problem is also able to solve TAHDS and hence, our problem is NP-hard. Therefore, we propose to use a heuristic algorithm to solve the problem because an extensive search of an optimal placement for example by means of backtracking is not feasible.

Adaptive Business Process Modeling in the Internet of Services (ABIS)

Monika Weidmann, Falko Koetter, Maximilien Kintz
 Fraunhofer Institute for Industrial Engineering IAO
 and University of Stuttgart IAT
 Stuttgart, Germany
 Email: *firstname.lastname@iao.fraunhofer.de*

Daniel Schleicher, Ralph Mietzner
 Institute of Architecture of Application Systems
 University of Stuttgart
 Stuttgart, Germany
 Email: *lastname@iaas.uni-stuttgart.de*

Abstract—In the Internet of Services many companies work together in interorganizational business processes. To enable ad-hoc business interaction it is necessary to align business processes of the business partners, especially in communication processes. These business processes can be partly standardized, but need to be slightly adapted for several similar use cases by the involved companies. This fosters adaptability and reuse for the business partners. We present an approach for adaptive business process modeling in the Internet of Services (ABIS), which allows creation of adaptable process templates. These templates are then used to create variants of processes allowing companies to work together in an interorganizational setting.

Keywords-business process management, adaptive business processes, internet of services, process modeling.

I. INTRODUCTION

The main idea of the Internet of Services (IoS) is to use the Internet as a medium for offering and selling services [1]. An infrastructure is needed to bring together service consumers and providers to trade services and enable the new business models, where organizations work together to deliver a service to consumers in a previously unknown manner [1] [2]. Business processes, which have been defined and owned by one company in the past, are now used to support a cross-company process flow [2]. In our work with insurance companies we experienced a need for standardized business processes, especially considering interorganizational communication processes, which provide a service, for example a repair service, for insurance customers. Although most companies wish for standardized reference processes to become available, there still persists a need for individualization. Additionally, the requirement to improve products, processes, and customer satisfaction, as well as changing market conditions, regulations, and laws cause a rising need for adaptation of business processes [3]. As in some business processes various partners are included [4], changes to the process affect the interorganizational communication directly. Companies are challenged to comply with different processes to communicate with their respective partners. Thus a need arises for multiple companies to adapt processes together, resulting in sound process models.

The main contribution of this paper is the introduction of concepts, which allow the creation of adaptable interorga-

nizational business processes based on a real use case and real business requirements. The goal is to enable business users with knowledge in process design to create process variants through direct interaction with the process model. In our approach called *Adaptive Business process modeling in the Internet of Services (ABIS)* we define new modeling elements to allow the creation of *process templates* in which *process fragments* may be inserted to create *process variants*. Process templates allow to model standardized and adaptive parts of a process including cross-company dependencies. Process fragments can be modeled independently by different participating companies. We call the process of creating a *variant* from such a template and fragments a *binding*. In contrast to the definition of the process templates and process fragments, binding the process template involves all participating parties. We chose BPMN 2.0 as notation because of its various abstraction levels and its increasing business support [4].

The remainder of this paper is structured as follows: In Section II, we analyze related work dealing with the adaptation of business processes and show the shortcomings we address in this paper. Section III describes a motivational example, which made apparent the need for adaptive business processes within this context. We use this as a continuous example throughout the paper. Section IV gives a detailed description of the introduced diagrams and modeling elements and shows how they can be applied to the use case. The future work is described in Section V before a conclusion is given in Section VI.

II. RELATED WORK

In this section we present related work with the focus on variable business processes. We compare selected approaches in a table according to different criteria before we introduce the variability model we use in our approach.

Previous work has been done on variability in software, for example in [5]. Recently, these concepts have been extended to provide variability in service-oriented systems [6], which combine services in order to provide higher level functionality (see also [7]). To compose services into a service chain, *executable process models* can be used [2]. Variability in process models can be added at *design time*

Table I
COMPARISON OF VARIABILITY MODELING APPROACHES

Criterion	Provop	PESOA	ProCon	MultiPers
Integrated variability visualization	-	✓	-	-
Responsibility modeling	-	-	(✓)	✓
Dependency modeling	✓	✓	✓	✓
Integrated dependency visualization	-	✓	-	-
WYSIWYG variant creation	(✓)	(✓)	-	-

and at *runtime*. We do not consider runtime variability as for example described in [8] or [9]. In our use case and project experience it is important to define a process model in advance as a guideline for business partners to be followed during the automatic or manual execution of the process. However, the runtime aspect of adapting business processes should not be disregarded in future work.

In the field of event-driven process chains (EPCs) much work has already been done towards configuration and adaptation [10] [11]. However, due to the complexity of the underlying approaches, and the missing direct interaction of the user with the process model, these approaches do not address the goals of ABIS.

We consider four most relevant concepts related to ABIS. The (1) *Provop* approach allows modeling of variability using so-called *options* on a basic process model, which alter the model by deletion, insertion, or modification operations [12] and has been extended with concepts to guarantee soundness [13]. (2) PESOA uses UML-like constructs for modeling *process families* [14]. A different approach called (3) *process configurator* (ProCon) allows explicit modeling of logic in a tree-based approach enhancing a process with variability [15]. A recent effort called (4) *Multi-Perspectives Variants* (MultiPers) defines a data structure to describe a family of process variants [16].

We compare the four approaches using the five criteria in Table I. As we plan to provide a multi-user approach, the modeling of responsibilities is needed. Next, dependencies between decisions are to be modeled and visualized within the process models for reasons of usability. Finally, we consider it important to provide a What You See Is What You Get (WYSIWYG) approach for the creation of variants, as in our experience many business users are already aware of process models [4]. However, a high usability is very important for business process management tools in general [4].

Different ways for modeling variable process models do not refer to one configurable process model, but the specification of single process fragments, which then can be glued together in order to reuse concepts and create different process models [17] and [18]. Although we will use concepts

of separate process building blocks, the need of the business users is a standardized process template to start with.

In our ABIS method we use the approach of Mietzner et al. in [19] and [7] for generic variability modeling in XML files. Here, the variability is added to XML files without altering the original file. This enables a separation of the process model (BPMN 2.0 XML file) from the variability (XML file) for storage and interchange of process models without extension of the BPMN 2.0 metamodel. The approach allows the definition of *variability points* and *alternatives* [19]. The alternatives can be explicitly defined, specified by the user, or be left empty. *Dependencies* allow to enforce a binding order of variability points, whereas *enabling conditions* can limit the choices for alternatives for variability points depending on previous choices. Based on this approach we can enable a multi-user derivation process for ABIS in future work, separate the variability from the process model, allow the modeling of complex dependencies, and provide tool support for the creation of variants supporting the user with automatic choices if only one alternative is left.

III. REAL WORLD USE CASE

In this section we present a simplified real world use case we came across on our work in the openXchange project (www.openxchange-project.de), dealing with creating a service network of small and medium sized enterprises to handle property damage claims. In active claims management, insurance companies often involve external partners for various tasks like creating a survey report or removing the damage. In the course of process standardization we came across individualization needs, as the companies want to:

- Work together with partners through IT-supported processes
- Use standardized predefined processes supporting their business needs
- Have individualization options for certain aspects of the business process
- Have a sound process model to communicate with their partners

We modeled a part of the active claims management process in Figure 1, where one simplified interorganizational business process is presented. The lanes in the process model were omitted to save space. At the top you see the customer's process, which in our example is an insurance company in need of a building repair service. Below it you can see the external partner or contractor. We reduced the detailed commissioning process to the following tasks: preparing, sending, and receiving requests and confirmations, and the handling of reports.

The adaptable parts for creating other process variants are highlighted in grey color and described using the text annotations (1), (2), and (3):

- At (1) it shall be possible to choose if intermediate reports are used. If they are used, the process looks like

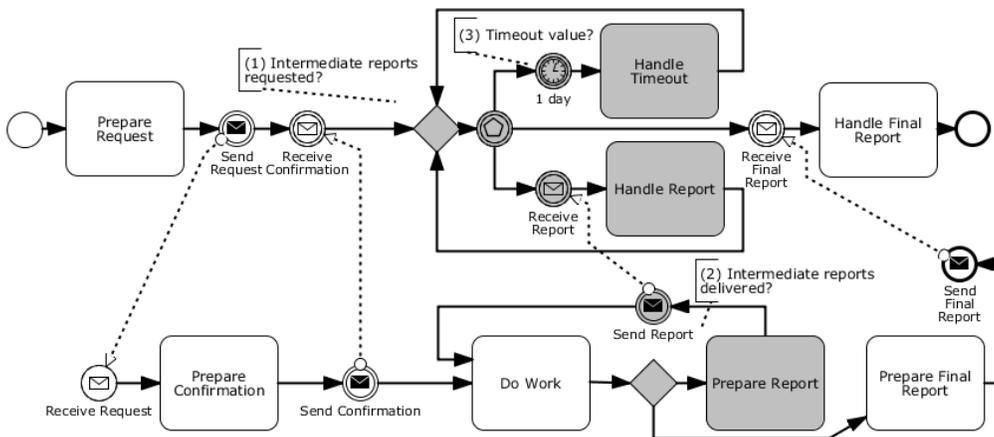


Figure 1. Real world use case: Active claims management

in Figure 1. Otherwise, the highlighted parts at (1) are omitted.

- At (2) it also shall be possible to configure the usage of intermediate reports. As it does not make sense to create reports which are not sent, (2) is considered *dependant* on the choice made at (1).
- At (3) the attribute *timeout* shall be configurable with the values '1 day' or '2 days'.

We will later show how these requirements are handled. Other possible use cases include the configuration of a B2B-online shop, the communication processes in case management (healthcare), or the configuration of air line catering processes.

IV. MODELING VARIABILITY IN BUSINESS PROCESSES WITH ABIS

In this section we show the scope of our variability modeling as a basis for the introduction of the ABIS concepts, which we then apply to the use case introduced above.

A. Variability scope

The list of BPMN 2.0 elements considered as variable so far is based on the BPMN 2.0 specification [20].

Event - There are multiple event types (error, escalation, etc.) defined in BPMN 2.0. They differ in the time dimension (interrupting vs. non-interrupting) as well as the flow dimension (starting a process, intermediate, ending a process). Changing these semantics via variability is error-prone from the business process designer view, as each of these events has to be used in a specific context. The resulting dependencies would add enormous complexity to a process model. We therefore do not consider the event types, time dimensions, and flow dimensions as variable to an event.

Activity - An activity can be a task or a sub-process (reusable and non-reusable). The choice of activities and sub-processes is one important element of variability in BPMN process models and is enabled by ABIS. Variability of activity

type will not be permitted explicitly, but may be modeled by offering a choice between different activities.

Gateway - Gateways allow steering the process flow. Therefore, the condition-expression of the gateway and the options are important places to add variability. Variability of the semantics of the gateway itself (exclusive, parallel, etc.) is difficult to understand for human process designers. The reason is that especially business users without technical knowledge find it hard to model detailed execution semantics [21], which are implied by several BPMN artifacts. Therefore, gateway types are not variable in ABIS.

Sequence and Message Flow - Variable sequence and message flow is a requirement for variable activities. If an activity is added to the process, the message flow will have to be adapted to connect the activity to the process correctly. Variable sequence and message flows are enabled by ABIS.

Pools and Lanes - The creation of variable pools and lanes implies a fundamental change of the business goal. Hence, it is not considered here.

Others - Other constructs are for example groups, text annotations, etc. These will not be considered specifically in this context but might be added in the future.

Attributes are considered separately. All mentioned elements have several attributes, for example name and documentation. From the use case, typically three types of attributes can be considered variable: gateway conditions, thresholds on sequence flows, and implementation details. As some attributes are defined for all elements, others only for a subset, it is difficult to restrict the attributes which can be considered variable. We generally consider all attributes to be variable.

B. Variability modeling elements in ABIS

Considering the variability scope defined in the previous subsection, the addition of variability can be done within the constructs of the chosen modeling language or by extending the modeling language with new constructs. Using extensions,

the visibility of variability within the process is higher than using native language elements. For tool support, it is also easier to use explicit variability modeling than implicit semantics. Because of these reasons, we introduce new constructs with the following goals:

- Introduce *as few and as simple constructs* as possible to ensure high usability and allow easy comprehension
- Provide an *additive approach*, as deletion is more expensive from a user’s point of view [22]
- Give a *graphical notation* which is not easily confusable with existing notations (considering BPMN 2.0)
- Enable the *scope*, as described in the previous section

To reach these goals, the following two *diagram types* are introduced:

Process template Within the process template, all BPMN elements are allowed. Additionally, elements for modeling variability are defined: *variable region*, *variable link* and *variable attribute*, which are then bound, resulting in a valid BPMN process variant.

Process fragments A process fragment in ABIS is a construct similar to a subprocess, which can be inserted into a process template at variable regions during *binding*. Process fragments in ABIS differ from subprocesses in two aspects. For one thing, they can define additional sequence and message flows to other elements of the process template. For another thing, they are inserted in the same scope as the variable region they replace. Process fragments may contain variable regions themselves. A process fragment is modeled with specialized start and end events called *fragment start and end links*. We do not allow deletion of process elements.

Process fragments could be either modeled within a process template or separately. We choose to model them separately, in order to fulfill the requirements of a distributed environment considering fragment repositories as has been researched in [17] and [18].

Additionally to these two diagram types, the following new *modeling elements* are introduced:

Variable region - A variable region is a new element for modeling similar to an activity. A variable region differs from an activity as follows: it has exactly one incoming and one outgoing sequence flow. A variable region is a placeholder for process fragments, which are inserted at this position in the current process template or process fragment.

Variable link - A variable link is a new element similar to the BPMN 2.0 throwing link event. In contrast to the throwing link event, a variable link is used to model the target of a message or sequence flow to show to which element this flow will be directly connected, that means without an additional catching link event. During the binding process, the incoming sequence or message flow of the variable link is connected to lead from its source to the specified *target* of the variable link. The link itself is then discarded.

Fragment start and end link - Fragment start and end links are used to denote where the incoming and outgoing

Table II
SCOPE - BPMN 2.0 ELEMENTS WITH VARIABILITY

BPMN Element	Can have variable attributes	Can be used in process fragments	ABIS variability modeling element
Event	✓	✓	
Activity	✓	✓	
Gateway	✓	✓	
Sequence and message flow	✓	✓	<i>variable link</i>
Pools and lanes	✓		
Others	✓	✓	

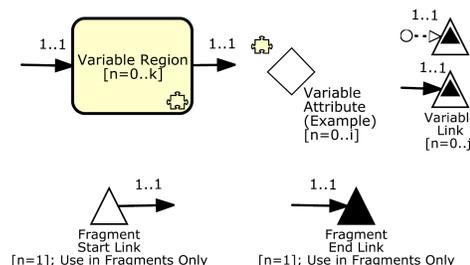


Figure 2. BPMN elements introduced

message flow needs to be connected in the processes template. Exactly one fragment start and end link have to be included in each process fragment.

Variable attributes - As all attributes may be variable, a separate description of the variability is needed.

Table II shows an overview of how variability is added to the BPMN 2.0 elements considered in the scope. The variability can be either added through variable attributes, by usage of the element in process fragments (and variable regions), or by using the explicit ABIS modeling elements.

We introduce a graphical notation for the new elements in Figure 2. The cardinality restrictions are shown in UML notation. The elements in the top row may be used as often as needed in process templates and process fragments. The elements in the bottom row are only allowed in process fragments. The notation of the variable attribute of the gate is an example, as all BPMN elements may have variable attributes. The reason for using puzzle pieces and triangles is that these shapes do not have semantics in BPMN.

The resulting BPMN model for process templates and variable regions will be stored as plain BPMN 2.0 with an additional XML file describing the variability points. Therefore, we do not define a BPMN 2.0 extension.

In the following section we will use the new constructs to show how the continuous example is adapted.

C. Application of ABIS to a real world use case

We have introduced an example for a commissioning process between a customer (insurance company) and a

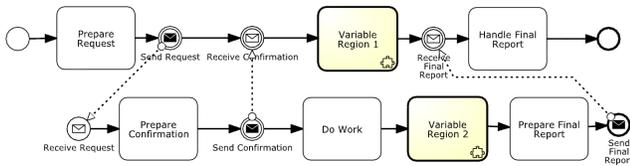


Figure 3. Process template for use case

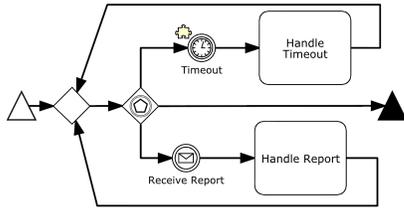


Figure 4. Process fragment A - option 1 for variable region 1

contractor (company offering repair services in buildings) in Section III in Figure 1. We take the following steps to model the variability:

- Identify the variable elements in the use case (already described in Section III)
- Create the process template for the use case (next step)
- Define the process fragments for the use case (see Figures 4 to 6)
- Model the structure of the use case (see Figure 7)

We will elaborate on a detailed methodology in future work.

Using the modeling elements and diagram types introduced before, we can replace (1) and (2) in Figure 1 by variable regions, as shown in Figure 3. Notice that the variable regions are marked with a puzzle piece. The variable timeout attribute has been placed in variable region 1. The variable regions have exactly one incoming and outgoing sequence flow.

The process fragment A in Figure 4 contains the first option for variable region 1. Here, the handling of the intermediate reports is modeled. Additionally, the timeout is variable and can be set to the different values a or b ('1 day' or '2 days'). Notice that all process fragments have exactly one fragment start and end link. Alternatively no intermediate reports are expected. Therefore, process fragment B - an empty fragment - is used (see Figure 5).

For the second variable region the first alternative is shown in Figure 6. It contains the preparation and the sending of intermediate reports. Notice the variable message flow to *Receive Report*. If no intermediate reports are requested, the empty fragment (see Figure 5) must also be used for variable region 2.

In order to bind the process template, the process fragments are inserted into the process template replacing the variable



Figure 5. Process fragment B - option 2 for variable regions 1 and 2

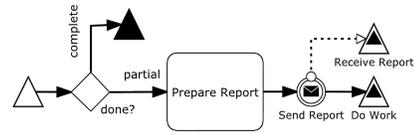


Figure 6. Process fragment C - option 1 for variable region 2

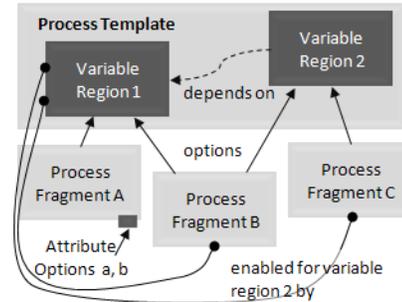


Figure 7. Structure of use case example

regions, resulting - when no variable elements are left - in a process variant. The sequence flows connecting to a variable region are connected to the respective start and end links. When the fragment has been inserted, all variable parts of the process fragment have to be bound. An example for this is the timer variable. The timeout value must be chosen after the insertion of process fragment A. For the example above, choosing process fragments A and C and setting the timeout to '1 day' would result in the initial process model (Figure 1).

In Figure 7 the *structure* of the example use case is shown. The two variable regions are marked in dark grey. The process fragments A, B, and C are the options for the variable regions as described above. Hereby, A and B can be *bound* to variable region 1, and B and C can be bound to variable region 2. Finally, the process fragment A also contains a variable attribute (the timeout), with the options a and b.

As variable region 2 is directly *dependant* on variable region 1, only one choice has to be made for variable region 1, which then directly affects variable region 2. The dependency between the variable regions 1 and 2 is indicated by a dotted line, showing that variable region 1 needs to be bound before variable region 2, as the variable link to *Receive Report* in variable region 2 (see Figure 6) would have no target otherwise. Additionally we use *enabling conditions* to limit the choices for variable region 2 according to the choice made in variable region 1. For formal definitions of the applied variability model see [7]. We will describe a detailed concept of how to apply the variability model to our modeling approach ABIS in future work.

V. FUTURE WORK

In future work we will develop a prototype which supports the creation of process templates, process fragments, and

binding alternatives to create a process variant. Additionally we will describe the architecture of the resulting solution, the process of filling in the variability points, as well as introduce an algorithm for handling of dependencies and automatic choice of alternatives based on previous decisions. We will validate our approach using a set of real world process models. In the course of this work we might consider defining additional variable BPMN 2.0 elements if needed. Further on we will work on a methodology with corresponding role models to support the application of our concept. We will consider combining the adaptable business processes with concepts of modifying runtime business processes like for example in [23]. However, these efforts strongly depend on the future development of BPMN 2.0 runtime engines. Finally we will analyze the possibility to enhance our approach with compliance-specific features in order to support design of compliant business processes as for example in [24].

VI. CONCLUSION

In this paper we introduced a method for adaptive business process modeling in the Internet of Services called ABIS. The goal of our approach is to enable business users to create their own process variants in an interorganizational setting based on standardized variable process models. To accomplish this, we introduce two new diagram types with additional modeling elements in BPMN 2.0. The business users may *separately* model parts of the process using *process fragments*, the first new diagram type. The second diagram type is the *process template*. Process fragments are inserted into process templates in order to create an interorganizational process variant. While fragments may be modeled independently, creating the process variant involves all participating parties. The presented concept was applied to a real world use case.

ACKNOWLEDGMENT

The work published in this article was partially funded by the openXchange project of the German Federal Ministry of Economy and Technology under the promotional reference 01MQ09011 and by the MASTER project (www.master-fp7.eu) under the EU 7th Research Framework Programme Information and Communication Technologies Objective (FP7-216917).

REFERENCES

- [1] J. Cardoso, K. Voigt, and M. Winkler, "Service engineering for the internet of services," in *Enterprise Information Systems*, ser. Lecture Notes in Business Information Processing, W. Aalst, J. Mylopoulos, N. M. Sadeh, M. J. Shaw, C. Szyperski, J. Filipe, and J. Cordeiro, Eds. Springer, 2009, vol. 19, pp. 15–27.
- [2] A. P. Barros and M. Dumas, "The Rise of Web Service Ecosystems," *IT Professional*, vol. 8, no. 5, pp. 31–37, Sep. 2006.
- [3] C. Wolf and P. Hermon, "The State of Business Process Management 2010," *BPTrends Reports (February 2010)*, 2010, last accessed 13.01.2011. [Online]. Available: www.bptrends.com/surveys_landing.cfm
- [4] S. Patig, V. Casanova-brito, B. Vögeli, C. Bern, and B. Voegeli, "IT Requirements of Business Process Management in Practice An Empirical Study," *8th International Conference on Business Process Management (BPM'2010)*, pp. 13–28, 2010.
- [5] I. Jacobson, M. Griss, and P. Jonsson, *Software Reuse: Architecture, Process and Organization for Business Success*. Addison-Wesley Professional, 1997.
- [6] S. H. Chang and S. D. Kim, "A variability modeling method for adaptable services in service-oriented computing," *11th International Software Product Line Conference (SPLC 2007)*, no. 10557, pp. 261–268, Sep. 2007.
- [7] R. Mietzner, "A Method and Implementation to Define and Provision Variable Composite Applications , and its Usage in Cloud Computing," Ph.D. dissertation, Universität Stuttgart, 2010.
- [8] M. Koning, C.-a. Sun, M. Sinnema, and P. Avgeriou, "VxBPEL: Supporting variability for Web services in BPEL," *Inf. Softw. Technol.*, vol. 51, no. 2, pp. 258–269, 2009.
- [9] C. Dorn, T. Burkhart, D. Werth, and S. Dustdar, "Self-adjusting Recommendations for People-Driven Ad-Hoc Processes," *8th International Conference on Business Process Management (BPM'2010)*, pp. 327–342, 2010.
- [10] M. Rosemann and W. M. P. V. D. Aalst, "A configurable reference modelling language," *Information Systems*, vol. 32, no. 1, pp. 1–23, 2007.
- [11] H. Reijers, R. Mans, and R. Vandertoorn, "Improved model management with aggregated business process models," *Data & Knowledge Engineering*, vol. 68, no. 2, pp. 221–243, 2009.
- [12] "Modellierung und Darstellung von Prozessvarianten in Propov," in *Modellierung 2008 Conference*, ser. LNI, T. Kühne, W. Reisig, and F. Steimann, Eds., vol. 127. GI, 2008, pp. 41–56.
- [13] A. Hallerbach, T. Bauer, and M. Reichert, "Guaranteeing Soundness of Configurable Process Variants in Propov," in *IEEE International Conference on E-Commerce Technology*, 2009, pp. 98–105.
- [14] A. Schnieders and F. Puhmann, "Variability Mechanisms in E-Business Process Families," in *Proc. International Conference on Business Information Systems (BIS 2006)*, 2006, pp. 583–601.
- [15] A. Werner and H. Müller, "Geschäftsprozesskonfigurator - Softwaregestützte Geschäftsprozessberatung," *ERP Management*, vol. 5, no. 2, pp. 25–28, 2009.
- [16] S. Meerkamm, "Configuration of Multi-Perspectives Variants," in *1st International Workshop on Reuse in Business Process Management (rBPM10)*, Hoboken, NJ, USA, 2010.
- [17] H. Eberle, T. Unger, and F. Leymann, "Process Fragments," in *On the Move to Meaningful Internet Systems: OTM 2009*, ser. Lecture Notes in Computer Science, R. Meersman, T. Dillon, and P. Herrero, Eds. Springer Berlin / Heidelberg, 2009, vol. 5870, pp. 398–405.
- [18] D. Schumm, D. Karastoyanova, F. Leymann, and S. Strauch, "Fragmento: Advanced Process Fragment Library," in *Proceedings of the 19th International Conference on Information Systems Development (ISD'10)*. Prague, Czech Republic: Springer, 2010, pp. 1–12.
- [19] R. Mietzner, A. Metzger, F. Leymann, and K. Pohl, "Variability modeling to support customization and deployment of multi-tenant-aware Software as a Service applications," in *PESOS '09: Proceedings of the 2009 ICSE Workshop on Principles of Engineering Service Oriented Systems*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 18–25.
- [20] Object Management Group (OMG), "Business Process Model and Notation (BPMN) Version 2.0," 2009, last accessed 13.01.2011. [Online]. Available: <http://www.omg.org/spec/BPMN/2.0/>
- [21] B. Silver, *BPMN Method and Style: A levels-based methodology for BPM process modeling and improvement using BPMN 2.0*. Cody-Cassidy Press, 2009.
- [22] O. Holschke, "Impact of Granularity on Adjustment Behavior in Adaptive Reuse of Business Process Models," in *Business Process Management*, 2010, pp. 112–127.
- [23] K. Vidačković, N. Weiner, H. Kett, and T. Renner, "Towards business-oriented monitoring and adaptation of distributed service-based applications from a process owners viewpoint," in *Service-Oriented Computing, IC3OC/ServiceWave 2009 Workshops*, ser. Lecture Notes in Computer Science, A. Dan, F. Gittler, and F. Toumani, Eds., vol. 6275. Springer Berlin / Heidelberg, 2010, pp. 385–394.
- [24] D. Schleicher, T. Anstett, F. Leymann, and R. Mietzner, "Maintaining compliance in customizable process models," *On the Move to Meaningful Internet Systems: OTM 2009*, pp. 60–75, 2009.

Context Factors for Situational Service Identification Methods

René Börner, Matthias Goeken

Process Lab, IT Governance Practice Network
Frankfurt School of Finance & Management
Frankfurt/Main, Germany
[r.boerner;m.goeken]@fs.de

Thomas Kohlborn, Axel Korthaus

Information Systems Discipline
Queensland University of Technology
Brisbane, Australia
[t.kohlborn;axel.korthaus]@qut.edu.au

Abstract—Service identification is one of the earliest and very crucial activities in a service engineering lifecycle and requires adequate methodological support in order to be successful. Although there are numerous service identification methods to be found in the literature, most of them take a one-size-fits-all approach that fails to acknowledge the broad variety of concrete circumstances that can form the organizational context in which these methods need to be applied. In this paper, we argue that there is a need for configurable service identification methods that can be tailored to their particular application contexts using situational method engineering. As a first step towards this goal, we analyze two explorative case studies and related literature to derive a basic set of relevant context factors that can influence and determine the final configuration of situational service identification methods from available method fragments. Adapting service identification methods to concrete project situations will improve their applicability and lead to a better service design.

Keywords - *Service-oriented architectures; service identification; service analysis and design; situational method engineering; context factors*

I. INTRODUCTION

Service orientation is a highly recognized paradigm in enterprise architecture. There are a number of expected benefits related to service-oriented architectures (SOA) in a technical and in a business-oriented sense. Although the business-oriented benefits, like flexibility, reusability and standardization, are of high importance [1], up to now, development of SOAs is mainly technically driven so that most approaches consider technical aspects in the first place [2].

The Organization for the Advancement of Structured Information Standards (OASIS) defines service-oriented architecture as a “paradigm for organizing and utilizing distributed capabilities that may be under the control of different ownership domains” [3]. Since the generic term “capabilities” can refer to both business functions and application functionalities, this definition supports a holistic SOA view that accommodates for two specific types of services: The term business service describes an autonomous, transformational capability that is offered to and consumed by external or internal customers for their benefit [4]. These services can have different levels of granularity ranging from comprehensive offerings (e.g., purchasing services) to fine-granular services (e.g., address

verifications) [5]. While flexibility and reusability usually increase when services become smaller, performance tends to deteriorate [6]. The second type, software services, enables a close business and IT alignment in order to support business services and thus the agility of organizations [7]. Software services expose application functionalities that can be re-used and composed based on business needs. In order to implement SOAs successfully, an adequate identification of these services is essential.

For the last couple of years many authors have been looking at the identification of services. A distinctive feature of identification approaches is the direction of the analysis. Some authors start from a Business Process Management perspective and follow a so-called top-down approach [8, 9]. Business processes are identified and subsequently broken down to activities. Finally, IT services are designed to support business functionality. In contrast, [10] start from a technical point of view and identify services bottom-up. Based on an asset analysis, e.g., the invocation frequency of certain applications can be analyzed to identify potential services. Usually, neither bottom-up nor top-down approaches are used in their pure form. Thus, many authors advocate hybrid service identification approaches that utilize techniques covering the analysis of both business processes and existing IT infrastructure [11, 12]. A comparison of further approaches can be found in [13].

Interestingly, most existing methods to identify services are based on a “one-size-fits-all” approach (for an overview see [14, 15]), i.e., they do not consider a configuration of methods depending on different circumstances such as the goals or various context factors of an SOA implementation. Even if context factors are considered, the scope of possible configurations is usually very limited [16]. Situational method engineering (SME) offers an opportunity to engineer service identification methods depending on situation-specific context factors of the project at hand. For this purpose, so-called method fragments are configured to methods that are adaptable to different situations.

The objective of this paper is to explore, which context factors affect the selection of method fragments for service identification and how they influence the development of situational methods. A qualitative analysis approach was chosen to analyze data from two case studies inspired by the constant comparative analysis method of grounded theory [17] in order to identify relevant context factors for service identification methods.

The remainder of this paper is structured as follows: In Section 2, scope and methodology of this paper are discussed. Section 3 describes the conducted case studies. The fourth Section will identify relevant context factors and their influence on fragment selection. Finally, Section 5 concludes the paper with a summary, current limitations and an outlook for further research.

II. SITUATIONAL METHOD ENGINEERING IN SERVICE IDENTIFICATION

This paper can be seen as part of a broader research program. The latter results in a comprehensive meta method to configure situational methods for service identification including the description of possible situations and available fragments. Following [18], we assume that “a richer understanding of a research topic will be gained by combining several methods together in a single piece of research or research program” (p. 241). Thus, in our research process, we combine the two research methods *case study research* and *design science research*. Within this context, the identification of context factors through case studies and the construction of a method supported by principles from SME can each be seen as separate research projects. Jointly, they are part of the research program, i.e., the development of a meta method for the configuration of situational methods for service identification (Fig. 1).

Despite the popularity of SOA, there is only little understanding of how to convey all advantages frequently mentioned in related literature. Moreover, little is known about how context factors impact service identification approaches in SOA. This corresponds to a low uptake of empirical research in systems and software development in general [19].

Against this background, we believe that qualitative case study research can make a useful contribution. Case studies are particularly relevant for research in its “early, formative stages” [20, 21] which applies to the field of SOA (see also [22] and [23]). As case studies can be descriptive and explorative in nature, they are supposed to give insights into how context factors influence service identification.

Service identification is one of the earliest activities in a service engineering process, which covers the whole lifecycle of a service. It is of particular importance, as any errors made during this activity can flow through to and build up in the design and implementation phases, which results in increased cost due to necessary rework [24]. A review of service analysis methods in general and service identification in particular by [14] reveals that none of the recently published methods is comprehensive and integrated enough to cover both SOA concepts (business and software services) to an adequate extent. However, as pointed out by the authors, different methods can complement each other and may have specific characteristics that make them more suitable in certain contexts.

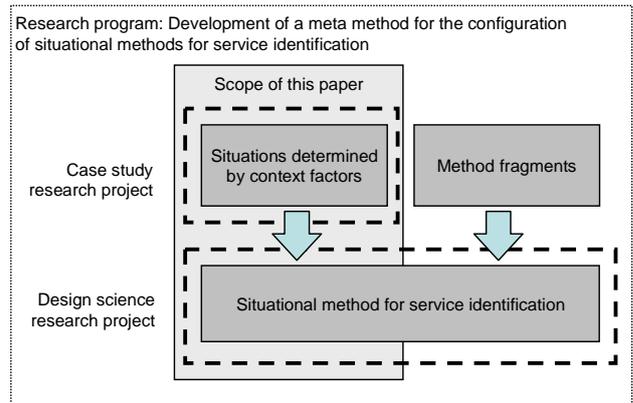


Figure 1. Research projects and research methods

For several years there have been efforts to guide the development of such methods in order to guarantee a high level of quality. To give this guidance is the task of method engineering (ME). ME is a discipline in information systems research meant to “design, construct and adapt methods (...) for systems development” [25]. The most popular approaches to ME [26-29] all identify activities, roles, results and techniques as important elements of methods [30].

Based on the fact that a given method m constructed at the time t_1 cannot fit all conceivable conditions and circumstances when it is used at a future time t_2 , the concept of situational method engineering emerged. The configuration of methods in SME is based on situations, i.e. once a situation is identified, a suitable method is configured. Reference [31] discusses how situations can be described satisfactorily. It concludes that context factors of concrete service identification projects are important for identifying situations in order to configure situational methods.

To provide for the configurability of a method, so-called fragments are constructed and afterwards configured depending on the situation [32, 33]. Reference [34] defines method fragments as “standard building blocks based on a coherent part of a method. A situational method can be constructed by combining a number of method fragments” (p.360). For the purpose of this paper, the notion of method fragment will be defined as any reasonable combination of method elements, i.e., activities, roles, results and techniques. The development of method fragments particularly for the purpose of service identification is part of our research program but out of the scope of this paper (Fig. 1). However, Table 1 will give indications on how the selection of appropriate method fragments is influenced by the context factors identified in Section 4.

Based on two case studies, this paper elaborates on relevant context factors for service identification projects and their influence on the configuration of situational methods. Identifying situations through context factors is thus an important pillar of a meta method that supports the design of methods for service identification.

III. CASE STUDY DESCRIPTIONS

The following case studies describe two SOA implementation projects conducted in Australian companies, namely Suncorp and the Securities Industry Research Centre of Asia-Pacific (SIRCA). The explorative nature of the case studies was meant to discover relevant context factors for the identification of services. At the same time, the significantly diverse settings of both cases opened up a continuum of instantiations for these context factors [35]. At Suncorp, researchers from the Queensland University of Technology (QUT) conducted an action research study. They actively participated in the project and helped test and apply a service analysis and design methodology developed by the researchers. In the second case study, several research methodologies were used. One of the most important sources of evidence have been interviews with SIRCA's employees and researchers from the University of New South Wales (UNSW), which were conducted shortly after the project had been completed. The interviews have been transcribed and analyzed afterwards.

A. Suncorp

In the context of an Australian ARC Linkage project titled "Service Ecosystems Management for Collaborative Process Improvement" (ARC Linkage Grant: LP0669244), some of the authors have developed a comprehensive service analysis and design (SAD) methodology for both business and software services. Thus, the SAD methodology used in this case study follows a hybrid approach to service identification. The first part covers the identification and analysis of business services by detailing, adapting, and consolidating existing service analysis approaches that focus on the business domain of an organization. This part is structured into four distinct phases, each comprising a specific set of activities that may use the outputs of previous phases as inputs. Subsequently, the second part of their approach describes how software services can be identified and analyzed that support business services in order to achieve close business and IT alignment. Similar to the first part of the consolidated approach, this part is structured into distinct phases, each comprising specific activities. A detailed description of the methodology is provided in [36]. The experiences gained through this exercise will build the foundation for the following discussion.

Suncorp is a diversified company in the financial services sector. As one of Australia's leaders in banking, insurance, investment and superannuation focusing on retail customers and small to medium businesses, the Suncorp Group is Australia's sixth largest bank and third largest insurer. The Suncorp case study can be subdivided into three phases as the organization went through different change programs related to their take on service orientation.

The first phase (1) focuses on service identification for integrating different systems. Suncorp had started a Claims Business Model Program some time ago with the intent to identify process improvements that would result in reduced leakage, reduced payments of ineligible claims, and lower handling costs. Suncorp's current systems were not flexible enough to support the required changes [37]. It was then

decided that the new claims process should be implemented in a new insurance claims management system from Guidewire Software, the ClaimCenter application [38]. While the initial implementation was in support of personal home claims, implementation projects for claims in worker's compensation, personal motor, commercial property and others followed. In the ClaimCenter project, integration with a large number of external systems was required and thirteen development teams (including the external vendor and an offshore team) using different development methods had to be coordinated. The project team decided to use an SOA approach to integrate diverse systems such as policy, payments, receipting and claims. A standardization of interfaces was sought to improve reusability [37].

Against this background, the QUT project team came in and presented the consolidated SAD methodology to a solution architecture team from Suncorp's Business Technology group, which started the second phase (2). Suncorp had found that their approach to SOA and service analysis and design was rather ad-hoc and very much driven by bottom-up integration requirements of their pilot projects, potentially lacking strong alignment of the service designs with the business processes:

"The current process that we follow tends to be driven by the functional requirements and data requirements of the consumer. This results in a very entity-driven service in which the consumer of the service needs to understand a lot more about the state and context of the call that they are making." (Suncorp Solution Architect)

A study protocol specified the objectives and the scope of the collaboration as well as the timeframe and the planned deliverables. In a first step, the AR study primarily focused on the identification of software services. The "motor claims" business process was chosen as input for the software service preparation and identification steps of the SAD methodology developed by QUT. To keep the scope manageable, two sub-processes, namely "claims intake" and "assessment" were selected. The researchers were provided with Suncorp's "motor claims" process models on different levels of hierarchy and additional business artifacts including the SOA Roadmap, the Insurance Domain Model and the ClaimCenter Hub System Architecture Specification. Based on this input, the researchers used the service preparation and identification steps prescribed by the SAD methodology for the two sub-processes and produced two reports that included the resulting service designs.

The third phase (3) with Suncorp extended the work that has been done previously, by applying the complete SAD methodology [36]. In particular, as part of a collaboration project between industry and university, three industry students applied the SAD methodology to derive business and software services starting from Suncorp's business strategy and capabilities. The project was driven by the desire to identify software services that not only support processes but also represent constituent elements of business services, which in turn needed to be identified first. Consulting QUT researchers along the project, the three students were able to apply the prescribed methodology and present their results to the business and IT audiences within

Suncorp who largely benefited from the lessons learned of this exercise. A detailed report was provided at the end of the project.

B. SIRCA

In the context of the project “Ad-hoc Data Grids Environments” (ADAGE), researchers at the UNSW implemented a service-oriented architecture for SIRCA. The project aimed at providing researchers an easier retrieval and analysis of heterogeneous data from different sources (grid environment) spontaneously in an unforeseeable fashion (ad-hoc). Neither business processes nor SOA were the core focus of the project. The former were hardly considered at all whereas the latter was chosen as the preferred architectural paradigm of this project. However, services (and their identification) were used as a means to meet SIRCA’s requirements rather than being the subject of analysis themselves.

In ADAGE, services were created based on the available data. This implied a technical understanding of services, which is also reflected by the synonymous use of the terms “service” and “web service” by project team members. Hence, the scope of service identification in this project was limited to software services.

SIRCA provides a huge data repository containing historical financial market data such as news and trading data. Their aim is to supply this data to researchers especially at Australian and New Zealand universities. Thus, their business model is fairly simple and is covered by one business process only.

“Our processes are fairly (...) atomistic. In that way, we are very simple outfit, we are a data repository, we collect lots of data, we do fairly substandard processing to it to normalize it and make it easily accessible. And then people access the data with some fairly straight-forward enterprise in that regard.” (Representative of SIRCA)

SIRCA employees were not thinking in terms of business processes, so that no model was delivered that could have been analyzed in the course of service identification. SIRCA’s management however had some requirements in mind that should be fulfilled by services. Unfortunately, these were not documented, which makes traceability difficult. Requirements were communicated to the project team in scheduled weekly meetings and workshops. Service candidates were identified on the basis of these meetings and prototyped. In an iterative and incremental approach the functionality of these candidates was adjusted to finally meet SIRCA’s requirements. In some cases, services were completely dismissed and new ones had to be created. A close collaboration between SIRCA’s research and development department and UNSW’s project team was a key to ensure the successful identification of services.

SIRCA’s management did not aim at the implementation of an SOA in particular. The idea of services was basically advocated by UNSW’s project team. Thus, there was no know-how on SIRCA’s side as far as SOAs are concerned. On the outset of the project, a funding for three years was provided. At the end of the project in December 2009, funds

for further six months were provided to implement the prototype and make it accessible to SIRCA’s customers.

First and foremost, the search for services was driven by the idea to retrieve and integrate data from different sources. In a second step, project members came up with ideas, which services could support researchers in analyzing data. This included, for example, building time series of financial data, merge data from different sources and visualize events. Only after that the identified and implemented services should be offered to third parties to support their business processes. Clearly, this was a requirements-driven bottom-up approach. Goals included the provision of a graphical user interface (GUI) to customers enabling them to directly invoke services in an ad-hoc fashion to analyze financial market data. This implies a distinct degree of customer interaction which influences the identification of services significantly. Certainly, SIRCA’s case is not a typical example of service identification projects. Because of its rather extreme character it helps to identify possible instances of context factors that usually cannot be found in typical cases.

IV. CONTEXT FACTORS IN SERVICE IDENTIFICATION

In order to engineer situational methods, relevant context factors that determine different categories of situations have to be identified. Hence, in this section we build on the case studies described previously to identify these context factors. The presentation of each context factor is structured as follows. Firstly, observations from the case studies are the basis for identifying context factors. Secondly, findings in related literature are briefly discussed where applicable to support the relevance of the encountered factors. Thirdly, an analysis of how these context factors influence the selection of fragments is conducted. Table 1 summarizes the results.

In SIRCA’s case, all services were clearly meant to be exposed to researchers from associated universities, i.e. to external **service consumers**. A graphical user interface (GUI) provides the opportunity for users to combine these services. Thus, users can choose themselves which services they need to use in order to analyze their data. In Suncorp’s first phase services were identified to integrate system functionalities. Hence, service consumers were purely internal. Functionalities that had to be accessed by using different applications were wrapped and can now be invoked as services. Conversely, the second and third phase aimed at both internal and external service consumers as certain services were intended to be used by end-consumers, internal systems and/or entities such as departments within Suncorp.

As shown in the case studies, services can be provided for different service consumers, e.g. other divisions (internally), third parties (externally) or both. Services can be used to integrate heterogeneous enterprise applications [9, 39] and simplify the access to certain functionalities for staff, i.e. for internal customers only. If this is known a priori, a number of activities and results such as the creation of an inter-organizational service map are not applicable in this situation. Moreover, there might be legal constraints that only apply if services are offered to third parties. Passing on customer data for instance must be permitted by the customer in some countries.

TABLE I. CONTEXT FACTORS AND THEIR INFLUENCE ON FRAGMENT SELECTION

Context Factor	Parameter Value	Influence on Fragment Selection
Service consumer	<ul style="list-style-type: none"> - internal - external - both 	<ul style="list-style-type: none"> - inter-organizational service maps are not applicable for purely internal services - "line of visibility" particularly important for external consumers - data flow analysis combined with legal check for externally provided services
Budget	<ul style="list-style-type: none"> - restrictive budget - generous funding 	<ul style="list-style-type: none"> - fragments supporting strategic aspects might not be used due to a restrictive budget - a generous funding enables comprehensive use of fragments - restrictive budgets often result in the selection of IT-oriented fragments that demand less resources
SOA concepts	<ul style="list-style-type: none"> - business services - software services - hierarchy of services 	<ul style="list-style-type: none"> - BPM techniques and respective fragments are essential for a business service-oriented identification - technical fragments such as asset analysis might be sufficient for the identification of software services - a comprehensive, hybrid approach using fragments with both top down and bottom up techniques is necessary for a hierarchy of services
SOA maturity level	<ul style="list-style-type: none"> - SIMM level 1-3 - SIMM level 4-7 	<ul style="list-style-type: none"> - fragments focused on SOA governance have only little meaning for maturity levels 1-3 - for SIMM levels 4-7 e.g. inter-organizational service maps can be important
Compliance	<ul style="list-style-type: none"> - general laws - industry-specific regulations - internal policies 	<ul style="list-style-type: none"> - all organizations have to ensure compliance with general laws such as consumer data privacy and must use respective fragments when service consumers are outside the company - a fragment that analyzes the necessity of industry-specific approaches should always be used - depending on the industry, additional fragments are needed - if there are internal policies, fragments that e.g. ensure consistent naming of services have to be applied
IT department	<ul style="list-style-type: none"> - existent - not existent 	<ul style="list-style-type: none"> - inputs like architectural concepts are only available from IT departments and restrict the use of method fragments - in the absence of an IT department, fragments are not applicable if they demand roles such as IT administrator
Interaction	<ul style="list-style-type: none"> - customer interaction - employee interaction 	<ul style="list-style-type: none"> - for the interaction with employees, roles prescribed by fragments have to be available - fragments delivering swim lane diagrams and analyzing the "line of visibility" are particularly useful if customer interaction is pivotal

Thus, fragments that demand an analysis of respective laws and regulations are only necessary where such data is passed on to third parties. An analysis of consumer interaction can be important for internal and external service provision. A fragment analyzing the "line of visibility" is much more important if services are exposed to external customers [11] and would add lots of value in cases like SIRCA's. However, they are less relevant in a context as given in Suncorp's first phase.

In both projects the **budget** seemed to play an important role and we perceived that budgeting has a significant impact when it comes to choosing necessary fragments of a method for service identification. Generally, a setting as encountered in SIRCA's case with a budget that allows for an extensive time frame of three years provides the opportunity for a thorough and systematic application of identification methods. You would expect the utilization of many techniques in order to ensure a high quality of implemented services. A detailed analysis of all available strategic and technical documents would be typical in such circumstances. However, in SIRCA's case the absence of such documents naturally dominated the generous time scope and made the application of many techniques impossible. Again, the head of research and development certainly devoted enough resources in SIRCA's case but still failed this broad analysis. The following citation is an excerpt of an interview conducted for the analysis of the ADAGE project.

"We didn't do the asset identification, because we didn't have anyone who knew about that, but if we had then, we would have done it. So you are constrained by the cost it

takes to develop, but also you were constrained by people's skills." (Researcher at the UNSW)

While the first project was funded internally, the second and third phases at Suncorp were basically developed in a collaborative environment between Suncorp and QUT based on mutual in-kind contributions. As such, the phases could be treated as pilots and were not associated with any costs other than research and development project budget for Suncorp. Budget information about the first phase is confidential and cannot be reported here. However, due to the partly academic character of these projects, funding restrictions were not a problem. At Suncorp, this availability of resources was used to apply QUT's identification method diligently.

Literature broadly confirms that the budget has implications on the number of available staff, the time pressure and the possibility to incorporate external help from consultants [40]. The higher the project sponsor's position in the company's hierarchy, the more likely is a generous funding. This allows for a proper analysis especially of strategic aspects and the inclusion of business processes and a comprehensive use of fragments. An initiation of a service identification project by the middle management (which is commonly accompanied by smaller budgets) often results in more pragmatic or technically-driven SOA implementations. Fragments dealing with Business Process Management as well as business process driven approaches (top down) are likely to be omitted in such cases. Instead, technically-oriented fragments analyzing applications and IT functionalities (bottom up) are used since they often promise quicker results. To make up for limited employee skills, a

larger budget is necessary because the company has to rely on external support. If funding is limited, certain fragments cannot be used due to this lack of skills.

Indeed, it is worth discussing if the budget is a factor as such and can be put on the same level as the other factors. In some cases, services might be identified without considering the budget available for their implementation. Only after the service identification process, a prioritization (depending on budget restrictions) is conducted [41]. However, we argue that fragments dealing, e.g., with value analyses are parts of the service identification method. These fragments are more likely to be selected if budget restrictions are tough which qualifies the budget as context factor.

The case studies showed clearly that different understandings of services based on different **SOA concepts** influence the proceeding of service identification significantly. In Suncorp's first phase aiming at the integration of different systems, the underlying technical SOA understanding implied a focus that lay purely on developing reusable software services. As part of the second phase – with a Business Process Management perspective in mind – software services that support business processes or at least sub-processes were the goal of the analysis. The third phase at Suncorp focused on identifying software services that support business services. Rarely, the focus on software services is as clear as in SIRCA's case.

A system integration approach like in Suncorp's first phase concentrated on software services, so that method fragments reflecting a more business-related SOA understanding and targeting the identification of business services were not considered. In the second phase, service identification had to include activities related to the analysis of process models and involved such fragments with not only IT-related, but also business-related staff roles. In Suncorp's third project both SOA concepts have been addressed. If the service identification is limited to a rather technical point of view like in SIRCA's case, the set of method fragments to be considered will be limited. Frequently, a hierarchy of services is the outcome of an identification process [42]. Higher-level services that support business processes compose finer-grained software services that adhere to the technical preconditions of underlying systems and data. In order to achieve this more complex type of an SOA, a broader range of method fragments has to be used complementarily. This is reflected in many hybrid (or "meet-in-the-middle") approaches that can be found in related literature [13, 15].

We observed different degrees of previous experience in the examined cases as far as SOA and service orientation is concerned. Based on this observation, we concluded that the **SOA maturity level** a company has achieved plays a role in the configuration of methods for service identification. SIRCA did not have any services at all when the project was initiated. At Suncorp, SOA maturity can be considered as relatively low. First projects had been conducted in the area of implementing software services. However, these projects did not aim at understanding business requirements and their impact on service-orientation but rather understanding the integration requirements of existing applications. It can be

conjectured that the maturity related to the adoption of service-oriented concepts will increase over time, which will change the scope of the SOA understanding and consequently the way service identification has to be conducted, as can be seen in the two latter Suncorp phases.

The SOA maturity level can be distinguished following the Service Integration Maturity Model (SIMM) by [43]. Levels 1-3 describe a rather low service orientation whereas companies with a SIMM level of 4-7 are more advanced in the field of SOA. If the latter is the case, an analysis of service maps among different divisions of a company can be essential to build an enterprise-wide SOA [44]. A fragment providing such an analysis cannot be sensibly used if there is no service orientation and subsequently (almost) no services. The same is true for all fragments dealing with SOA governance and inter-organizational aspects of SOA. If the SIMM level is low, fragments that deliver, e.g., inter-organizational service maps cannot reasonably be applied. Thus, the selection of appropriate fragments is influenced by the SIMM level of a company.

Compliance issues can arise from both legal obligations and regulatory restrictions as well as from internal policies and may require corresponding service identification method fragments that address these issues. In SIRCA's case no particular regulatory or legal requirements had to be considered. However, the Reuters market data provided by SIRCA must not be used by everyone. It might only be used for academic purposes. The academic institution has to pay a subscription fee to SIRCA to give their employees access to the data. Hence, restricted access to data and the intended use of services must be considered when services are identified. As far as Suncorp is concerned, for example the general insurance reform act and related laws and regulations issued by the Treasury Department of the Australian Government are industry-specific requirements. Since implemented services in Suncorp's case are exposed to customers, confidential treatment of sensitive customer data had to be guaranteed by a proper service design. Furthermore, as already indicated, Suncorp generally follows an agile approach to developing services. Thus, methods related to the identification, design and implementation of services have to comply with the agile paradigm.

In case services are provided to third parties, a fragment that guarantees customer data privacy and security has to be applied. Therefore, interactions with and data flows towards all service consumers have to be analyzed. Certain industries such as banking, insurance or pharmaceuticals have to adhere to additional, stricter regulations and should use respective fragments. Internal policies – such as restrictions on software development methods – can make some fragments inapplicable. Other fragments may be necessary to fulfil for instance internal naming conventions.

In a small company that lacks an **IT department** (like SIRCA), methods have to be adapted to accommodate for this circumstance. Responsibility for IT is commonly distributed all over the company and departments tend to implement isolated IT solutions or so-called silos. Larger organizations like Suncorp are usually structured along the lines of business but have an IT division that takes care of a

company-wide IT architecture and infrastructure. Hence, the existence of a designated IT department and thus the degree of centralization of the IT infrastructure is an important context variable.

On the one hand, a high degree of centralization or the existence of a central division supervising and governing IT implementation throughout a company usually leads to more transparency. Frequently, at least some information on applications and data is readily available. This can be used as input for service identification method fragments. On the other hand, some fragments demand certain roles such IT administrators or newly designed units consisting of business and IT employees (see also [45]). In a small company that lacks an IT department, these method fragments are often not applicable. The company size (frequently considered to have an important influence on SOA implementations [46]) and the geographic scope of operations are thus closely linked to the existence of a central IT department and subsequently not considered as context factors themselves.

Due to the fact that in all cases participation, exchange and contribution of service consumers differed notably, we hypothesized that varying degrees and forms of **interaction** with both customers and employees necessitate the use of different method fragments. In SIRCA's case, for instance, employees are not directly involved in service delivery because the services are very fine-grained and fully automated. The coarser-grained services are, the greater is the possibility that they are only semi-automated or manual and subsequently interact with employees. Customer interaction is of high importance in SIRCA's case because the ad-hoc composition of services is a primary goal.

However, due to their fine-grained nature, services themselves are executed independently from users, i.e., no customer interferes directly in a service. At Suncorp, employees were only involved to showcase the developed methodology and gain information about current practices at Suncorp. As part of the third project, different employees at Suncorp were involved to identify business services.

Getting access to and involvement of business roles was difficult in Suncorp's case but required by the used method fragments. Where this is impossible, a different method configuration is necessary. In general, a customer interaction can be obligatory in some places or can happen "on demand" if required by the service or desired by the customer [47]. In automated services possible customer interaction has to be foreseen and planned for. If customer interaction is a major issue for the identification of services in a situation at hand, related method fragments are crucial for a successful implementation. One example are swim lane diagrams that show interfaces to customers.

All context factors described above were found in the two case studies. Their effect on the selection of method fragments for service identification is summarized in Table 1. Particularly the case study at SIRCA revealed some more potential context factors. Since their relevance could not be observed in the phases at Suncorp, these factors were omitted from the discussion in this paper. Probably, there are interdependencies and relationships between the identified context factors. Analyzing these relationships and identifying

relevant combinations that determine the situational configuration of methods is out of the scope of this paper and will be the focus of future work.

V. CONCLUSION

Not only literature but also experience shows that methods in information systems research should be configurable depending on the situation at hand, i.e., in a situation-specific way. This is also true for service identification methods in service-oriented architectures. In order to support a situation-specific configuration of such methods, situations have to be defined by context factors. The latter determine which method fragments should be used in the course of an identification process. Based on qualitative research using grounded theory, the data of two case studies was analyzed in this paper to identify seven context factors, compare them with existing literature and describe how actual instances of these factors can influence the selection of method fragments. Due to the explorative nature of the two case studies, there is no guarantee that the derived list of factors is comprehensive. Further case studies might reveal other relevant factors. Moreover, the relevance of the identified factors cannot be proven by our case studies. Only the application of a complete situational method to a service identification project could attest their relevance.

As a prerequisite for situational methods, method fragments that can be combined have to be created. Thus, either parts of existing methods have to be identified as feasible method fragments or new fragments have to be created [48]. Designing these fragments to meet the requirements of situation-specific service identification is left to future research. Therefore, it is necessary to investigate which activities, techniques, roles, results and sequences can be combined reasonably. The so designed method fragments have to be configured to suit concrete situations that can be characterized by the context factors identified in this paper. Hereby, interdependencies and influences among context factors have to be analyzed. As mentioned previously, a concrete instantiation of one context factor can dominate other factors in the selection of one fragment over another. When exactly this is the case or how to weigh situational factors to come to a best possible selection of fragments must be elaborated in more detail.

Finally, a comprehensive situational method for service identification needs to be developed. It should define how to configure existing method fragments depending on the situation at hand. The herein identified context factors are critical to identify these situations. In order to evaluate the quality of so configured methods, more case studies should be carried out. In contrast to the two explorative case studies used to derive situational factors, further case studies should apply a newly created situational method to prove its applicability and evaluate its concept.

ACKNOWLEDGMENT

Parts of this research have been funded by a research project within the Australian Research Council Linkage Schema (grant code LP0669244), including financial support from SAP Research and the Queensland Government.

REFERENCES

- [1] M. Durst and M. Daum, "Erfolgsfaktoren serviceorientierter Architekturen," *HMD - Praxis der Wirtschaftsinformatik*, vol. 44, no. 253, 2007, pp. 18-27.
- [2] E.G. Nadhan, "Seven Steps to a Service-Oriented Evolution," *Business Integration Journal*, no. 1, 2004, pp. 41-44.
- [3] OASIS, "Reference Model for Service Oriented Architecture 1.0," <http://docs.oasis-open.org/soa/v1.0/soa-rm.pdf>, 2006.
- [4] M. Rosemann, E. Fieft, T. Kohlborn, and A. Korhau, *Business Service Management*, in: *Smart Services CRC White Paper no 1: 2009*.
- [5] A. Sehmi and B. Schwegler, "Service-oriented modeling for connected systems - part 1," *The Architecture Journal*, vol. 7, 2006, pp. 33-41.
- [6] J. den Haan, "SOA and Service Identification," <http://www.theenterprisearchitect.eu/archive/2007/04/26/soa-and-service-identification>, 2007.
- [7] L. Cherbakov, G. Galambos, R. Harishankar, S. Kalyana, and G. Rackham, "Impact of service orientation at the business level," *IBM Systems Journal*, vol. 44, no. 4, 2005, pp. 653-668.
- [8] V. Winkler, "Identifikation und Gestaltung von Services. Vorgehen und beispielhafte Anwendung im Finanzdienstleistungsbereich," *Wirtschaftsinformatik*, vol. 49, no. 4, 2007, pp. 257-266.
- [9] A. Arsanjani, S. Ghosh, A. Allam, T. Abdollah, S. Ganapathy, and K. Holley, "SOMA: A method for developing service-oriented solutions," *IBM Systems Journal*, vol. 47, no. 3, 2008, pp. 377-396.
- [10] T. Böhm and H. Krcmar, "Modularisierung: Grundlagen und Anwendung bei IT-Dienstleistungen." in *Konzepte für das Service Engineering. Modularisierung, Prozessgestaltung und Produktivitätsmanagement*, T. Herrmann, U. Kleinbeck, and H. Krcmar, Eds., Heidelberg: Physica, 2005, pp. 45-83.
- [11] K. Klose, R. Knackstedt, and D. Beverungen, "Identification of Services - A Stakeholder-Based Approach to SOA Development and Its Application in the Area of Production Planning," *Proc. 15th European Conference on Information Systems (ECIS)*, 2007, pp. 1802-1814.
- [12] F. Kohlmann and R. Alt, "Business-Driven Service Modeling - A Methodological Approach from the Finance Industry," *Proc. Business Process & Service Computing: First International Working Conference on Business Process and Services Computing*, 2007, pp. 180-193.
- [13] D. Birkmeier, S. Glöckner, and S. Overhage, "A Survey of Service Identification Approaches," *Enterprise Modelling and Information Systems Architectures*, vol. 4, no. 2, 2009, pp. 20-36.
- [14] T. Kohlborn, A. Korhau, T. Chan, and M. Rosemann, "Service Analysis - A Critical Assessment of the State of the Art," *Proc. 17th European Conference on Information Systems*, 2009.
- [15] R. Börner and M. Goeken, "Methods for Service Identification," *Proc. 7th International Workshop on Modelling, Simulation, Verification and Validation of Enterprise Information Systems (MSVVEIS)*, 2009, pp. 76-84.
- [16] R. Börner, "Applying Situational Method Engineering to the Development of Service Identification Methods," *Proc. 16th Americas Conference on Information Systems*, 2010, Paper 18.
- [17] B.G. Glaser and A.L. Strauss, "The Discovery of Grounded Theory: Strategies for Qualitative Research," Chicago: Aldine Publishing Company, 1967.
- [18] J. Mingers, "Combining IS Research Methods: Towards a Pluralist Methodology," *Information Systems Research*, vol. 12, no. 3, 2001, pp. 240-259.
- [19] M. Jarke, "Perspectives in the Interplay Between Business and Information Systems Engineering and Computer Science," *Business & Information Systems Engineering*, vol. 1, no. 1, 2009, pp. 70-74.
- [20] I. Benbasat, D.K. Goldstein, and M. Mead, "The case research strategy in studies of information systems," *MIS Quarterly*, no. 11, 1987, pp. 269-386.
- [21] M.D. Myers, "Qualitative Research in Business & Management," London: Sage Publications, 2009.
- [22] H. Luthria and F.A. Rabhi, "Building the Business Case for SOA: A Study of the Business Drivers for Technology Infrastructure Supporting Financial Service Institutions." in *Enterprise Applications and Services in the Finance Industry*, D. Kundisch, D.J. Veit, T. Weitzel, and C. Weinhardt, Eds.: Springer, 2009, pp. 94-107.
- [23] R.A. Stebbins, "Exploratory Research in the Social Sciences," Thousand Oaks: Sage Publications, 2001.
- [24] S. Inaganti and G.K. Behara, "Service Identification - BPM and SOA Handshake," *BPTrends*, vol. 3, 2007, pp. 1-12.
- [25] J. Ralyté, S. Brinkkemper, and B. Henderson-Sellers, "Situational Method Engineering: Fundamentals and Experiences." in *Situational Method Engineering: Fundamentals and Experiences*, J. Ralyté, S. Brinkkemper, and B. Henderson-Sellers, Eds., Boston: Springer, 2007, pp. V-VI.
- [26] T.A. Gutzwiller, "Das CC-RIM-Referenzmodell für den Entwurf von betrieblichen, transaktionsorientierten Informationssystemen," Heidelberg: Physica, 1994.
- [27] M. Heym, *Methoden-Engineering - Spezifikation und Integration von Entwicklungsmethoden für Informationssysteme*. St. Gallen: Institut für Wirtschaftsinformatik, Universität St. Gallen, 1993.
- [28] S. Brinkkemper, "Method engineering: engineering of information systems development methods and tools," *Information & Software Technology*, vol. 38, no. 4, 1996, pp. 275-280.
- [29] F. Karlsson, *Meta-Method for Method Configuration - A Rational Unified Process Case*. Linköping: Linköping University, Faculty of Arts and Sciences, 2002.
- [30] M. Goeken, "Entwicklung von Data-Warehouse-Systemen. Anforderungsmanagement, Modellierung, Implementierung," Wiesbaden: Deutscher Universitäts-Verlag, 2006.
- [31] T. Bucher, M. Klesse, S. Kurpjuweit, and R. Winter, "Situational Method Engineering - On the Differentiation of "Context" and "Project Type"." in *Situational Method Engineering: Fundamentals and Experiences*, J. Ralyté, S. Brinkkemper, and B. Henderson-Sellers, Eds., Boston: Springer, 2007, pp. 33-48.
- [32] N. Arni-Bloch and J. Ralyté, "Service-Oriented Information Systems Engineering," *Proc. 20th Conference on Advanced Information Systems Engineering (CAISE)*, 2008, pp. 140-143.
- [33] J. Ralyté and C. Rolland, "An Approach for Method Engineering," *Proc. 20th International Conference on Conceptual Modelling*, 2001, pp. 471-484.
- [34] P.J. Agerfalk, S. Brinkkemper, C. Gonzalez-Perez, B. Henderson-Sellers, F. Karlsson, S. Kelly, and J. Ralyté, "Modularization Constructs in Method Engineering: Towards Common Ground?" in *Situational Method Engineering: Fundamentals and Experiences*, J. Ralyté, S. Brinkkemper, and B. Henderson-Sellers, Eds., Boston: Springer, 2007, pp. 359-368.
- [35] R.K. Yin, "Case Study Research - Design and Methods," Thousand Oaks London New Delhi: SAGE Publications, 2003.
- [36] T. Kohlborn, A. Korhau, T. Chan, and M. Rosemann, "Identification and Analysis of Business and Software Services - A Consolidated Approach," *IEEE Transactions on Services Computing*, vol. 2, no. 1, 2009, pp. 50-64.
- [37] J.A. Couzens, "Implementing an Enterprise System at Suncorp Using Agile Development," *Proc. 20th Australian Software Engineering Conference* 2009.
- [38] Guidewire, "Guidewire Claimcenter - the Flexible Claims Management System for Property & Casualties Insurers," http://www.guidewire.com/our_solutions/claimcenter, 2009.
- [39] T. Erl, "Service-Oriented Architecture - A Field Guide to Integrating XML and Web Services," Upper Saddle River, NJ: Prentice Hall, 2004.
- [40] J. Becker, R. Knackstedt, D. Pfeiffer, and C. Janiesch, "Configurative Method Engineering," *Proc. 13th Americas Conference on Information Systems (AMCIS)*, 2007, Paper 56.
- [41] G. Stewart and A. Chakraborty, "Service Identification through Value Chain Analysis and Prioritization," *Proc. 16th Americas Conference on Information Systems*, 2010, Paper 70.
- [42] N. Josuttis, "SOA in der Praxis - System-Design für verteilte Geschäftsprozesse," Heidelberg: dpunkt.verlag, 2008.
- [43] A. Arsanjani and K. Holley, "The Service Integration Maturity Model: Achieving Flexibility in the Transformation to SOA," *Proc. IEEE International Conference on Services Computing (SCC'06)*, 2006, pp. 515.
- [44] F. Kohlmann and R. Alt, "Aligning Service Maps - A Methodological Approach From the Financial Industry," *Proc. 42nd Hawaii International Conference on System Sciences*, 2009, pp. 1-10.
- [45] R. Börner, S. Looso, and M. Goeken, "Towards an Operationalisation of Governance and Strategy for Service Identification and Design," *Proc. 13th IEEE International EDOC Conference*, 2009, pp. 180-188.
- [46] D. Sedera, "Does Size Matter? Enterprise System Performance in Small, Medium and Large Organizations," *Proc. 2nd Workshop on 3rd Generation Enterprise Resource Planning Systems*, 2008.
- [47] M. Leyer and J. Moormann, "Facilitating operational control of business services: A method for analysing and structuring customer integration," *Proc. 21st Australasian Conference on Information Systems*, 2010, Paper 42.
- [48] B. Henderson-Sellers and J. Ralyté, "Situational Method Engineering: State-of-the-Art Review," *Journal of Universal Computer Science*, vol. 16, no. 3, 2010, pp. 424-478.

Contract-Performing Circumstance-Driven Self-Adaptation and Self-Evolution for Service Cooperation

Ji Gao^{1,2}, Hexin Lv¹

¹ College of Information Science & Technology
Zhejiang Shureng University, Hangzhou, China, 310015

² College of Computer Science & technology
Zhejiang University, Hangzhou, China, 310027
gaoji1@zju.edu.cn hexin10241024@sina.com

Abstract—Service cooperation-based Virtual Organizations (VOs) have become the mainstream approach for developing application software systems in Internet computing environments. However, the large-scale deployment of VOs encounters serious difficulty due to the non-autonomy for their organization and maintenance. This paper focuses on VO self-maintenance and proposes the framework for achieving the contract-performing circumstance-driven self-adaptation and self-evolution of service cooperation, in order to maintain effectively the capability for a VO to achieve its objectives in two stages: self-adaptation and self-evolution.

Keywords—contract-performing; circumstance-driven; self-adaptation; self-evolution; virtual organization

I. INTRODUCTION

Constructing Virtual Organizations (VOs) by creating service cooperation (i.e., service-oriented cooperation) has become the mainstream approach for reforming the development of application software systems in Internet computing environments due to the development of Service-Oriented Computing (SOC) [1] and Service-Oriented Architecture (SOA) [2]. However, the current techniques for service cooperation are confronted with severe limitation: service cooperation is non-autonomic, making it unable to agilely adapt to the dynamically changing and unpredictable Internet cooperation environments. The leading cause is the inherent non-controllability of business services across different management domains (i.e., the consumer of a service can't control the process of the service provision). It is the non-controllability that brings on the so-called "trust" crisis that the success and benefit of cooperation cannot be ensured, and therefore makes service cooperation have to depend on a great deal of manual intervention.

Evidently, without the self-organization and self-maintenance of service cooperation, it is difficult to realize the large-scale deployment of VOs. Therefore, we have developed a series of research work for overcoming "trust" crisis and achieving autonomic service cooperation in the support of the National Science Foundation and the National High-Technology research and Development Program (863) of China. We have established a model oriented to multiagent systems, called IGTASC (Institution-Governed Trusted and Autonomic Service Cooperation) [3], to overcome "trust" crisis first. Then, based on IGTASC, we

have developed two frameworks to support the self-organization of VOs [4] and the self-maintenance of VOs respectively.

This paper focuses on the framework for achieving the self-maintenance of service cooperation, called CCAE (Contract-performing Circumstance-driven self-Adaptation and self-Evolution for service cooperation). The next section introduces the relative work and our countermeasure, including the foundation created by IGTASC. Then, Section III specifies the proposed framework CCAE in general. Section IV, V, and VI describe main constituents of CCAE: contract-performing circumstance model, Joint Contract-Conforming Mechanism, and VO Self-Adaptation and Self-Evolution Mechanism respectively. After the implementation and application analysis in Section VII, the conclusions and future work (in Section VIII) are provided.

II. RELATED WORK AND OUR COUNTERMEASURE

How to achieve the self-adaptation and self-evolution in abnormal situations is a difficult problem, worrying MAS (Multi-Agent System) researches for a long time [5][6].

2.1 Related Work

The current research for this problem is focused on the large-scale MASes composed of simple homogeneous agents, such as computing intelligence (evolution computing [7], artificial immunity systems [8], adaptive learning [9], etc.) and swarm intelligence [10]. However, the same research for small-scale MASes dynamically composed of self-interested, often much more complicated, heterogeneous agents (denoted by d-si-h-MASes hereafter) is much less and no systematic research results with practical value have been reported though such MASes are much more valuable and have the potential for large-scale deployment (see Section 2.2).

The main cause is that the methodologies of statistics, randomization, and optimization suiting computing intelligence and collective intelligence cannot be used in d-si-h-MASes, and again, there is no enough motivation and requirement for driving the researches adapting to d-si-h-MASes due to two hindrances. One is the inherent non-controllability mentioned above while the other is that the

MAS technology itself is disjointed with real-life application software systems [5]. Although there have been some self-healing research work (which belongs to self-evolution research category) for statically deployed small-scale MASes [11][12], the research results cannot adapt to open and dynamically configured d-si-h-MASes.

2.2 Our Countermeasure

The model of IGTASC mentioned above and its infrastructure can be used to conquer the two hindrances, and thereby create a substantial basis for researching the self-organization, self-adaptation, and self-evolution of service cooperation.

IGTASC proposes a three-level Virtual Society as the environment where VOs live and work (Figure 1): Agent Community, TAVOs (Trusted and Autonomic VOs), and Rational Agents, and depends on three technologies to make service cooperation both trusted and autonomic: Institution-Governed cooperation, Policy-Driven self-management, and Cooperation Facilitation management [3]. Also, reforming MAS technology by adopting the “service-oriented” concept removes the “disjoined” hindrance.

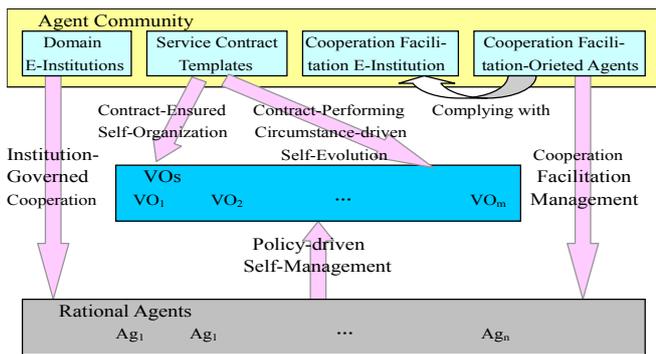


Figure 1 Three closely coupled mechanisms constituting IGTASC

IGTASC restricts the organizational form of a VO to the most familiar and widely-used cooperation form in human society: an alliance based on service providing-requiring relations, which is sponsored and created by some physical organization to satisfy a business requirement dynamically occurring (such as making new products, solving complex problems, searching for knowledge, purchasing merchandise, etc.). Such an alliance often concerns multiple binary collaborations which are managed by the sponsor centrally, but there are no interactions between other members (these interactions can be removed by partitioning business activities reasonably and arranging the appropriate messages sent by the alliance manager). Of course, every member of a VO should set up an agent as its broker for providing business services, and this makes the VO a typical d-si-h-MAS.

Because VOs are organized dynamically on user requirements (i.e., the newest objectives and tasks), and such VOs are of short life: the life-period of a VO ends once relevant user requirements are satisfied, this paper does not

consider the change of user requirements in a life-period, and only focuses on responding the abnormal change of cooperation circumstances.

The sponsor (and manager) of a VO should create and sign a providing-requiring contract with the provider of each outer service. Since these contracts specify, by contract-performing norms, the detail of bi-cooperation activities, the run of the VO becomes the interaction process for its members to complete cooperation according to contracts.

The social facilitation e-institution for cooperation facilitation management formulates not only the cooperation facilitation-oriented services, but also macro-level cooperation behavior norms, such as the obligations for service providing and requiring parties to comply with micro-level contract-performing norms and report the post-states for executing those norms. It is the execution of those macro-level norms that supports the in-time creation of contract-performing circumstances, which constitutes the foundation for achieving VO self-adaptation and self-evolution.

Based on IGTASC and the above countermeasure, we have proposed the framework of CCAE. By developing technologies of contract-performing circumstance model, joint contract-conforming mechanism, abnormal circumstance-driven VO maintenance, and 3-phase control cycle for transacting abnormal circumstances, CCAE can maintain the capability for a VO to achieve its established objectives effectively.

III. SELF-ADAPTATION AND SELF-EVOLUTION FRAMEWORK CCAE

CCAЕ supports the maintenance of run-time VOs (which are in d-si-h-MAS form) in two stages: self-adaptation and self-evolution. The former does not change the constituents of a VO, including its members and business process, while the latter requires replacing some members or even the business process. However, both stages depend on the mechanism of joint contract conformity driven by contract-performing circumstances.

The work model of CCAE is illustrated in Figure 2. It consists of joint contract-conforming mechanism, VO self-adaptation and self-evolution mechanism, contract-performing circumstance model, contract-performing circumstances, and service contract set. The joint contract-conforming mechanism manages contract-performing processes and monitors contract-performing circumstances. The management function enables the provider and consumer of a business service to execute in turn protocol entries in a service contract (say *sc1*), and creates in time, according to contract-performing circumstance model, the contract-performing circumstance (say *CPC^{sc1}*) to make both parties in service cooperation have a whole view of cooperation states. It is the whole view that drives the alternate and compact execution of contract entries and lets the monitoring function discovery in time the occurrence of contract violation events.

Once receiving a contract violation event, the VO self-adaptation and self-evolution mechanism activates the model ACVOM (Abnormal Circumstance-driven VO self-Maintenance), which drives a 3-phase control cycle to transact the abnormal circumstance indicated by the event. The VO self-Maintenance tries the self-adaptation first, and, if it fails, then tries the self-evolution.

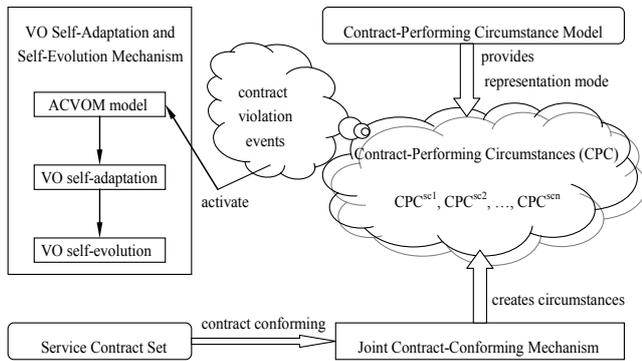


Figure 2 The work model of CCAE

IV. CONTRACT-PERFORMING CIRCUMSTANCE MODEL

This representation model describes the performing circumstances of service contracts. A typical Service Contract for business service bs , denoted by SC^{bs} , can be defined as a 3-tuple:

$$SC^{bs} = (BI, QoSG, CPP)$$

- BI: the Basic Information of service cooperation, which is used to specify the identity of both parties, the business transaction roles enacted by both parties, period of validity for this contract, service content (e.g., the operations or product items, price, number, deadline), payment mode, etc.

- QoSG: the QoS (Quality of Service) Guarantee, which defines quality parameters and metrics, and stipulates service level objectives (SLOs) based on those definition.

- CPP: the Contract Performing Protocol, which is designed as a partial-order set composed of protocol entries represented as contract-performing norms and uses BI and QoSG as the content referenced when executing those norms.

Definition 1 (contract-performing norm, cpn): define, with extended Deontic Logic [13], $cpn = OB_a^{sc} (\rho \leq \delta \mid \sigma) \mid FB_a^{sc} (\rho \leq \delta \mid \sigma) \mid PB_a^{sc} (\rho \leq \delta \mid \sigma)$, indicating respectively that, when σ holds true, party a (the agent signing SC^{bs}) is obligated to, forbidden to, or authorized to make ρ true before deadline δ (here ρ , δ , and σ are all the propositions describing contract-performing circumstances). Note, σ and ρ is often as the pre-condition and post-condition for executing cpn respectively.

Definition 2 (post-state of an executed cpn , ps): define $ps = (cpn\text{-number}, status\text{-type}, status\text{-description})$, where $cpn\text{-number}$ indicates the serial number of cpn in CPP of SC^{bs} , $status\text{-type}$ indicates the type of ps (success, fail, or exception), and $status\text{-description}$ gives the description of ps .

Next, an example of cpn is given, which comes from a supposed application of data mining.

```
(eio:Norm //A simplified norm for ensuring operator quality
NormNo: 21; //Norm 21 in contract performing protocol
Performer:"RespondingRole"; //The norm should be executed by provider
Trigger: (@eio:OperationCall Operator:"Mining" CallTime:?x);
//Triggered by an operator invocation event ("@eio:OperationCall
Operator:"Mining" CallTime <...>")
Deadline: (@eio:dateTimePeriod BeginTime:?x Period:?nego_03);
//The deadline completing the norm execution is ?nego_03 which begins
//from ?x. Herer, the value of ?x come from unification examination
//when this norm is triggered, and the value of ?nego_03 depends on the
//nigotiation between providing and requiring parties.
Postcondition: (@eio:SLOSatisfactionStatus SLOName:"SLOOfMining"
Operator:"Mining" Status:"True"); )
```

This cpn specifies: when the operator "Mining" provided by business service "DataMining" is invoked, an event indicating the invocation should occur and activate the cpn ; and, as the post-condition (i.e., $cpn.p$), the service level objective (SLO), named as "SLOOfMining", should be satisfied (i.e., the SLOSatisfactionStatus is "True") after cpn is executed. Thus, if this cpn is executed successfully, $cpn.ps.status\text{-description}$ (the status-description in ps of the cpn) must be of the same pattern as $cpn.p$ in order to match with $cpn.p$. That is, the status-description should be: (@eio:SLOSatisfactionStatus SLOName:"SLOOfMining" Operator:"Mining" Status:"True").

Based on the two definitions above, the Contract-Performing Circumstance for SC^{bs} , denoted by CPC^{sc} , is described with the sequence of $pses$:

$$CPC^{sc} = \{ps_1, ps_2, \dots, ps_m\}, ps_i = \text{post-state } (\mathcal{A}cpn_i), cpn_j (\in CPP \text{ of } SC^{bs}),$$

where the relevant $cpns$ are executed on the order specified by CPP of SC^{bs} , i^{th} post-state is denoted by ps_i , and $\mathcal{A}cpn_j$ indicates j^{th} norm in CPP is executed.

CPP divides $cpns$ into two types: backbone and compensation. While backbone $cpns$ indicate the main activities that must be executed for achieving the objective stipulated when the VO is sponsored, compensation $cpns$ are only used to recover the normal execution of contracts from abnormal $pses$ of executed $cpns$.

Definition 3 (abnormal ps of a executed cpn , aps): a ps is the aps iff $cpn.ps.status\text{-description}$ does not match with $cpn.p$.

Definition 4 (activation event for a cpn , ae): define ae as the part of ps of a executed cpn : $ae = (cpn\text{-number}, status\text{-type} \mid status\text{-description}, \text{ where } status\text{-type} = \text{'success'} \mid \text{'fail'} \mid \text{'exception'} \text{ and } status\text{-description} = \text{'(@<prefix>: <concept-name> \{<parameter-value>\}^*)'}$. (Note, @ indicates the instance of a concept defined in the domain ontology denoted by <prefix>.)

It is aes that drive the ordered execution of service contracts (i.e., $cpns$ included in them). While an ae coming from the ps of a successfully executed cpn activates the backbone cpn executed next, an ae coming from the aps activates one or more compensation $cpns$. Also, the abnormal change of contract-performing circumstances

indicated by *apses* drives the self-adaptation and self-evolution of a VO.

V. JOINT CONTRACT-CONFORMING MECHANISM

This mechanism is driven by contract-performing circumstances. The two macro-level cooperation behavior norms of obligation formulated in the social facilitation e-institution become the basis for implementing this mechanism: both service providing and requiring parties must comply with *cpns* and report *ps*es of executed *cpns* to each other. Also, the cooperation-facilitating service “ContractExecutionReport” for performing “reporting” obligation should be defined in this e-institution and configured to both parties.

The Joint Contract-Conforming Mechanism (JCCM) is represented as a multi-tuple:

JCCM = {vm-set, contract-set, cpn-set, self-executing, self-examining, inter-reporting, inter-examining}, where

- vm-set: the set of members in a VO;
- contract-set: the set of service contracts in the VO;
- cpn-set: the union of *cpn* sets; $\text{cpn-set} = \text{cpn-set}^{\text{sc}1} \cup \text{cpn-set}^{\text{sc}2}, \dots, \cup \text{cpn-set}^{\text{sc}n}$, where $\text{cpn-set}^{\text{sc}i}$ indicates the set of *cpns* formulated in CPP of contract sc_i (\in contract-set);

- self-executing: $\text{vm-set} \times \text{contract-set} \rightarrow \mathbb{P}\text{cpn-set}$; here, every vm (\in vm-set), according to the service contract sc (\in contract-set) signed by it, executes the *cpns* relevant to its obligations and rights; (hereafter, \mathbb{P} denotes power set and \rightsquigarrow denotes partial function.)

- self-examining: $\text{vm-set} \times \text{contract-set} \rightarrow \mathbb{P}\text{cpn-set}$, here, every vm (\in vm-set), according to the sc , examines the *ps*es of *cpns* executed by itself, and self-examining $(vm, sc) = \text{self-executing}(vm, sc)$;

- inter-reporting: $\text{vm-set} \times \text{contract-set} \rightarrow \mathbb{P}\text{cpn-set}$, here, every vm (\in vm-set), according to the sc , reports the *ps*es of *cpns* executed by itself to the opposing party of cooperation, and inter-reporting $(vm, sc) = \text{self-executing}(vm, sc)$;

- inter-examining: $\text{vm-set} \times \text{contract-set} \rightarrow \mathbb{P}\text{cpn-set}$, here, every vm (\in vm-set), according to the sc , examines the *ps*es of *cpns* executed by the opposing party of cooperation, and inter-examining $(vm, sc) \cup \text{self-examining}(vm, sc) = \text{cpn-set}^{\text{sc}}$, inter-examining $(vm, sc) \cap \text{self-examining}(vm, sc) = \emptyset$.

JCCM is installed into every business operation-oriented agent to implement two main functions: managing contract-performing processes and monitoring contract-performing circumstances. CPP of each contract sc (\in contract-set) becomes the basis for agent to manage the execution of sc . The contract-performing circumstance for sc changes continually along with the execution of *cpns* (\subset cpn-set^{sc}). If cpn_i (\in cpn-set^{sc}) needs to be executed by the agent itself, this agent should invoke the local operator relevant to cpn_i

before deadline δ , create *ps* according to the operation result, and report *ps* to the opposing party of cooperation.

Monitoring contract-performing circumstances includes the self-examining and inter-examining for *ps*es of executed *cpns*. The examinations are focused on whether or not a *ps* can be generated before the deadline and satisfy *cpn.p*.

It is the mutual reporting of *ps*es that enables both parties of each contract sc (\in contract-set) to observe and examine the whole contract-performing circumstance in time and thus to drive the alternate and compact execution of *cpns*.

Reporting actively *apses* (abnormal *ps*es) occurring in one's own side can facilitate the discovery and transaction of abnormality. Because the compensation *cpns* can be executed as soon as possible, this enhances the reliability and robustness of service cooperation.

VI. VO SELF-ADAPTATION AND SELF-EVOLUTION MECHANISM

This mechanism uses the model of ACVOM (Abnormal Circumstance-driven VO self-Maintenance) as the basis for implementing VO self-maintenance, and adopts a 3-phase control cycle as the framework.

6.1 Model ACVOM

ACVOM enables the VO sponsor, depending on its management policies, to centrally manage and control abnormal circumstance-driven VO maintenance. The maintenance activities are flexible and scalable: from the small ones such as modifying a service contract to the large such as replacing a service provider or even a business process.

The model of ACVOM is defined as a multi-tuple:

ACVOM = (CPC, CMP, cv-events, cm-principles, cm-plans, cm-actions, analysing, planning, executing), where

- CPC: the set of service Contract-Performing Circumstances; here, $\text{CPC} = \{\text{CPC}^{\text{sc}1}, \text{CPC}^{\text{sc}2}, \dots, \text{CPC}^{\text{sc}n}\}$, and $\text{CPC}^{\text{sc}i}$ indicates the circumstance of sc_i (i^{th} service contract);

- CMP: the set of management Policies which the VO manager (sponsor) possesses for supporting Cooperation Maintenance; here, $\text{CMP} = \text{an-policies} \cup \text{pl-policies} \cup \text{ex-policies}$, where an-policies, pl-policies, and ex-policies indicate the subset of analysis, planning, and execution policies respectively;

- cv-events: the set of contract violation events (*aes* from *apses*) reflecting CPC abnormality;

- cm-principles: the set of cooperation modification principles which are the result of contract violation analysis;

- cm-plans: the set of cooperation modification plans which are the planning result;

- cm-actions: the set of cooperation modification actions specified by cooperation modification plans;

- analysing: $\text{an-policies} \times \text{cv-events} \times \text{CPC} \rightarrow \text{cm-principles}$; here, the analysis activities denoted by this function adopt a domain-specific circumstance abnormality

analysis policy *anp* (\in an-policies), activated by *cve* (\in cv-events), to analyse CPC^{sc} (\in CPC) creating *cve* and propose an analysis result (a cooperation modification principle) *cmpr* (\in cm-principles);

- planning: pl-policies \times cm-principles \rightarrow cm-plans; here, the planning activities denoted by this function adopt a domain-specific cooperation modification planning policy *plp* (\in pl-policies), activated by *cmpr*, to drive planning and generate a cooperation modification plan *cmpl* (\in cm-plans);

- executing: ex-policies \times cm-plans \rightarrow Pcm-actions; here, execution activities denoted by this function adopt a domain-specific plan execution policy *exp* (\in ex-policies), activated by *cmpl*, to start cooperation modification actions (specified by *cmpl*) *cmas* (\subset cm-actions).

The above mapping functions of analysing, planning, and executing constitute jointly the 3-phase control cycle for transacting abnormal circumstances, and the activities in those phases are driven by CMP (Figure 3). Next, we only explain the transaction made by the VO sponsor. In fact, the transaction made by other VO members is similar and simpler.

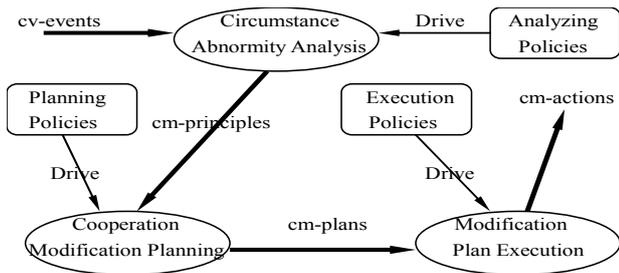


Figure 3 The policy-driven 3-phase control cycle for abnormal circumstance-driven VO maintenance

1) Circumstance abnormality analysis

The analysis work is driven by analysis policies. Once a *cve* (\in cv-events) occurs, the relevant analysis policy is triggered, which is used to analyse the cause, property, and effect of *cve* and select one from multiple activated compensation *cpns*. It is the multiple compensation *cpns* that enable the benefit-losing party to have multiple choices for maintaining contract execution. The selection depends on the integrated analysis of multiple factors, such as the respective results expected when executing these *cpns*, the work situation of current service contract-performing protocol, the whole execution situation of service cooperation in the VO, the business objectives of the VO sponsor, application domain knowledge, etc.

2) Cooperation modification planning

This phase aims at using planning policies to generate relevant cooperation modification plans in order to eliminate the minus affect of abnormality or to decrease the affect to the least degree. The extent of cooperation modification depends on the effect of execution of compensation *cpns*,

whether there are spare service providers or not, the structure of current business process, etc. and therefore can be partitioned into the next types (from the small to the large):

- replace the provider of a service (because the contract for this service is violated);
- defer in turn the time for providing following services in the current business process;
- replace the current business process with new one, including to cancel the contracts for all services not occurring in the new and to create the contracts for new services occurring in the new one;
- dismiss the VO and cancel the contracts for all services in the current business process.

Based on planning policies, extent-different modification plans can be created, and hence make the adaptation and evolution of cooperation display the better flexibility and scalability.

3) Modification plan execution

This phase aims at applying execution policies to detail modification plans and execute modification actions. For example, a modification plan only indicates to replace the provider of a service while the activities for determining the new provider of this service, making negotiation, signing the contract with this new provider, etc. should be driven by execution policies.

Evidently, it is the proper transaction of abnormal circumstances that supports the self-adaptation and self-evolution effectively.

6.2 VO self-adaptation and self-evolution

We view both self-adaptation and self-evolution of a VO as the means to maintain the capability for the VO to achieve its own objectives (Figure 4). The main difference is the extent of VO change: the former does not change VO constituents while the latter requires replacing some VO members or even the business process. In fact, the former can be viewed as the first stage for responding abnormality, and only when it does not bear fruit, the VO maintenance enters into the second stage: self-evolution.

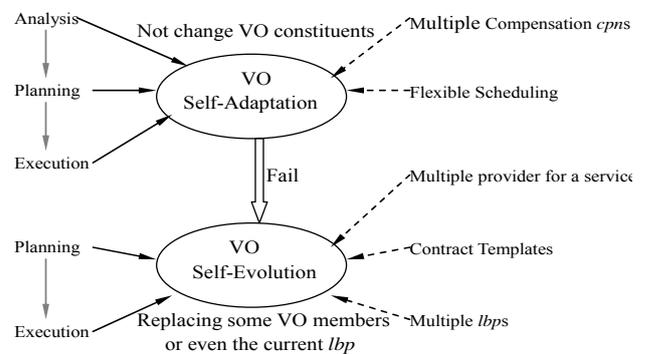


Figure 4 VO self-adaptation and self-evolution

1) VO self-adaptation

This stage depends on two key technologies: the flexible scheduling of the local business processes (denoted by *lbps* hereafter) and the configuration of compensation *cpns* in CPP (contract-performing protocol) of SC^{bs}. Here, *lbps* are formulated as the activity-scheduling plans for an agent to achieve local business objectives.

Although there are a variety of domain-specific abnormal circumstances, the mode for transacting them is the same: create a requirement for selecting suitable one from the compensation *cpns* and use the requirement as an event to activate the policy starting a 3-phase control cycle.

Next, we explain, by the supposed application of data mining, the policy-driven self-adaptation based on a flexible scheduling method. Figure 5 illustrates two *lbps* for achieving a data mining task. Each *lbp* displays only the activities, indicated by circles, performed by invoking the outer business services. Suppose the execution of activity 2 (in *lbp* 1 of Figure 5) generates an abnormal circumstance because the outer service *bs* for performing this activity is unavailable before the deadline. In order to transact the violation, two compensation *cpns* have been configured in CPP of SC^{bs}: number 01 and 02, which can be activated at the same time. The former informs *bs* provider to make *bs* available before a later deadline while the latter cancels the contract for *bs*.

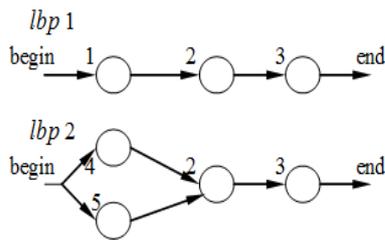


Figure 5 The *lbps* for a data-mining task

The policy for selecting from the two *cpns* is formulated as following:

```

Policy //The policy for selecting from cpn 01 and 02
Name: "CircumstanceAbnormityAnalysisForServiceAvailability";
PolicyType: "Obligation" ;
Processing: (ruleGroup ATFSA); //The processing work after this
//policy are activated: make a choice by executing rulegroup ATFSA
Target: (@Service "Monitoring" "GS");
Trigger: (@ContractNormConflictOccur ActivatedNormNo1:?x
ActivatedNormNo2:?y) ($= ?x 01) ($= ?y 02); //The
//trigger condition: cpn 01 and 02 are activated at the same time
End Policy
ruleGroup ATFSA
mode: p; // Denotes the production reasoning
select: first; //Execute the first activated rule
ruleList:
(→ ($ServiceAvailablePeriodAnalysis)
($BBSStore ($CreateConceptInstance "SelectedNorm" 01)));
(→ ($ExtendedAvailablePeriodAnalysis)
($BBSStore ($CreateConceptInstance "SelectedNorm" 01)));
($SendMessage "Monitor" ($CreateConceptInstance
"PlanningDeferringOfFollowingServices" 01 02));
End ATFSA
    
```

This policy is activated by the *cpn* selection requirement, and then it starts, by executing rule-group ATFSA, the analysis phase of a 3-phase control cycle to analyse the availability of *cpn* 01. The analysis work includes:

- Check whether a later deadline for *bs* can be assigned by executing the Boolean function as the condition part of rule 1 "ServiceAvailablePeriodAnalysis". If it can, the function returns 'true', and further results in the execution of *cpn* 01.

- Or else, by executing the Boolean function as the condition part of rule 2 "ExtendedAvailablePeriodAnalysis", calculate whether a extended later deadline for *bs* can be assigned. If it can, the function returns 'true' and drives the execution of *cpn* 01.

- Or else, send an internal message "@PlanningDeferringOfFollowingServices 01 02" to the agent (VO manager) itself by executing rule 3 (the unconditional rule).

This message activates the policy for driving the planning of service deferring. Then this policy starts the planning phase to determine which in the following services *bses* to be deferred (e.g., *bs* for performing activity 3 in Figure 5) and the deterred time for them. If the deferring plan is generated, a policy for driving the execution of this plan is activated to start the deferring negotiation with the providers of *bses* (the details are omitted).

In summary, it is the agent management policies that drive effectively the self-adaptation of service cooperation and VOs. Especially, the flexible scheduling of *lbps* and the configuration of compensation *cpns* not only enhance the possibility for removing abnormal circumstances but also facilitate the survival of current *lbps* in abnormal circumstances. Therefore, this enables VOs to agilely respond abnormity due to not changing VO members and *lbps*.

2) VO self-evolution

If the contract for *bs* must be cancelled (by executing *cpn* 02), the effort for maintaining VO capability enters into the second stage: self-evolution. Because the circumstance abnormity analysis has already been done in self-adaptation stage, only the latter two phases of a control cycle: Cooperation modification planning and modification plan execution, need to be performed.

The work for VO evolution planning is as following:

- Determine whether a new provider of *bs* can start *bs* in the available or extended period of *bs*. If can, send an internal message "@ExecutingReplacementPlan "ReplacingServiceProvider" <*bs*> <new provider>" to the agent (VO manager) itself in order to drive the replacement of *bs* provider.

- Or else, evaluate a later deadline and drive the replacement of *bs* provider.

- If the replacement fails, send an internal message "@Executing ReplacementPlan "RplacingBusiness Process"" to the agent itself in order to drive the replacement of the current *lbp* with a new *lbp*.

- Or else (a new provider is found), send a message to activate the policy for driving the planning of deferring (similar to the planning phase of VO self-adaptation).

If a new provider is found and the deferring plan is generated, the policy for plan execution phase is activated.

The plan execution activities include to negotiate the service provision with new provider and to negotiate the service deferring with the providers of following services. Once all of these negotiations are completed successfully, *lbp* 1 of Figure 5 can be resumed.

When the above planning or plan execution fails, replacing this *lbp* is necessary if there are spare *lbps*. Therefore, the planning and execution work for replacing the current *lbp* with a new *lbp* should be driven by relevant policies, including: find an available *lbp* (e.g., *lbp* 2 in Figure 5), cancel the contracts for *bs* relevant to activity 1, negotiate, create and sign the contracts for business services relevant to activity 4 and 5, and finally perform *lbp* 2.

If no new *lbp* is available, or, for any one from *bse*s relevant to activity 4 and 5, no applicable provider is found, the VO must be disbanded, and all the contracts for other *bse*s in *lbp* 1 must also be cancelled. Of course some compensation work must be done by executing compensation *cpns* formulated in those contracts.

VII. IMPLEMENTATION AND APPLICATION ANALYSIS

We have already created the self-adaptation and self-evolution framework CCAE by intensifying the agent platform included in the infrastructure for IGTASC.

First, a monitor module is nested into the platform to perform CPP of SC^{bs} depending on the contract-performing circumstance model and joint contract-conforming mechanism and to implement the activities for monitoring abnormal circumstances and the ones in the 3-phase control cycle. Because the platform has provided the policy engine, it is easy to make the work of monitor module become policy-driven as long as configuring a set of domain-specific policies and the operators and functions driven by these policies.

Second, the policy-driven self-management enables business operation-oriented agents to rationally conform to macro-level cooperation behavior norms formulated in the social facilitation e-institution, especially the obligations for complying with micro-level *cpns* in service contracts and reporting the post-states of executed *cpns*. Besides, the uniform facilitation service “ContractExecutionReport” configured to those agents creates the basis for implementing the joint contract-conforming mechanism and monitoring abnormal circumstances.

Third, formulating one or more service contract templates for each business service simplifies the creation, negotiation, and conformation of contracts. It is those templates that enable application domains to formulate statically parameterized *cpns*, and thereby enable business operation-oriented agents to possess, by statically configuring operators specified by *cpns* and management policies, the capability for executing *cpns*, achieve flexible scheduling of *lbps*, and implement policy-driven self-adaptation and self-evolution.

We have established several experimental service cooperation-based VOs, such as small meeting arrangement, knowledge provision, data mining, multi-part device

cooperation production, and multi-department crisis cooperation transaction, and used those VOs to test and validate the self-adaptation and self-evolution of service cooperation and VOs. The experimental results indicate that CCAE can support the run and maintenance of VOs effectively in very different application domains.

Next, we adopt the supposed application of data mining (see figure 5) to make an analysis. The current VO dynamically created wants to complete a data-mining task by invoking the three sequential business services provided by different partners (see *lbp* 1 in Figure 5). Because the VO sponsor and these partners all are the rational agents supported by CCAE and registering in the agent community, this VO can implement the joint contract conformation, transact abnormality, and maintain VO capability effectively.

In order to validate the compact and fluent execution of the data-mining task in normal circumstances, we let the three services for performing the three activities in *lbp* 1 to be assigned equal-length available periods first, and then make those periods overlap with each other. The process for executing this task indicates that the three services can be provided one after another with no time interval between them.

In another test, the service 1 for performing activity 1 is not available before the deadline. This contract violation event activates compensation *cpns*: 01 and 02, and further activates policy “CircumstanceAbnormalityAnalysisForServiceAvailability” (see Section 6.2). Because the provider can make service 1 available in a later deadline (which does not affect the provision of sequential services), this policy drives the execution of *cpn* 01: specify the new deadline, and inform the provider.

Again, if the provider can not make service 1 available in the later deadline, the message “(@PlanningDeferringOfFollowingServices 01 02)” is sent to activate the policy for driving the planning of service deferring. The produced deferring plan indicates that the provision of service 1 and 2 should be deferred in order to enable service 1 to be provided before the later deadline. And then the policy for driving the execution of deferring plan is activated to drive the deferring negotiation with the providers of service 1 and 2. Due to the success of negotiation, this policy drives the execution of *cpn* 01 before the later deadline.

Evidently, It is the flexible scheduling of *lbps* and the configuration of multiple compensation *cpns* that enable the VO to achieve self-adaption without changing its constituents.

We also validate the self-evolution of this VO by testing two instances. One is the fail of deferring negotiation while the other is the QoS violation in providing service 1. Both instances bring on the cancel of the contract for service 1, and thereby this drives the process for finding new provider of service 1. We examine two situations: finding one or no new provider. The former results in the update of single VO member while the latter brings on the replacement of *lbp* 1 with *lbp* 2 (see Figure 5), and the bigger change of VO

constituents: removing the provider of service 1 and increasing the providers of service 4 and 5.

Here, the basis for implementing VO self-evolution is the standards for business services and business operation-oriented roles and the coupling cooperation behavior norms formulated in e-institutions, and the formulation of multiple compensation *cpns* while the policy-driven cooperation management enables the general agent platform to be specialized into adapting to application domain requirement by configuring domain-specific policies and policy-driven operations and functions.

In summary, the advantage of CCAE is induced as follows:

1) Realizing the self-maintenance of VOs in d-si-h-MAS form; this enables those VOs not only to be composed dynamically and on requirement depending on the model IGTASC and its infrastructure, but also to maintain their capability for achieving objectives effectively, and thereby creates the solid foundation for the large-scale deployment of VOs

2) Since the cooperation between VO members focuses, by using contracts as tie, the monitoring of cooperation circumstance on the execution states of service contract-performing protocols, this makes the discovery of cooperation exceptions and the self-maintenance for eliminating exception impact have a reliable and accurate basis.

3) Configuring to every business operation-oriented agent the uniform facilitation service “ContractExecution Report” and the obligation for reporting the post-states of executed *cpns* enables both provider and consumer of a service to acquire in time the whole service contract-performing circumstance, and accordingly facilitates the compact execution of contract-performing protocols and the discovery of contract violation exceptions.

4) The policy-driven self-adaptation and self-evolution not only enables business operation-oriented agents to rationally conform to macro-level cooperation behavior norms formulated in the social facilitation e-institution, but also makes, by formulating domain-specific policies, the general model CCAE specialized easily into adapting to different application domains.

5) Dividing the maintenance of run-time VOs into two stages (self-adaptation and self-evolution) enables service cooperation not only to gain compact and fluent execution by self-adaptation (due to no change of the constituents of a VO), but also to adapt to, by self-evolution, the complex situation requiring replacing some members or even the business process.

VIII. CONCLUSION AND FUTURE WORK

This paper focuses on the self-maintenance of VOs in d-si-h-MAS form, which are much more valuable and have the potential for large-scale deployment, and has created the framework CCAE to achieve the contract-performing

circumstance-driven self-adaptation and self-evolution for service cooperation. CCAE can maintain effectively the capability for a VO to achieve its dynamically established objectives in two stages: self-adaptation and self-evolution, and thereby enhance the survival of VOs and current business processes for scheduling service cooperation.

The future work will be the formalization of CCAE and the development of real-life application systems based on CCAE.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the National Science Foundation of China (Grant 60775029), the National High-Technology research and Development Program (863) of China (Grant 2007AA01Z187), the Priority Theme Emphases Project of Zhejiang Province, China (Grant 2010C11045), and the Natural Science Funds of Zhejiang Province, China (Grant Y107446).

REFERENCES

- [1] M. P. Papazoglou, P. Traverso, S. Dustdar, et al. Service-oriented computing: state of the art and research challenges. *IEEE Computer*, 40 (11): 38-45, 2007.
- [2] M. Stal. Using architectural patterns and blueprints for service-oriented architecture. *IEEE Software*, 23(2): 54-61, 2006.
- [3] J. Gao, H. Lü, H. Guo, et al. Trusted autonomic service cooperation model and application development framework. *Science in China Series F: Information Sciences*, 52 (9): 1550-1577, 2009.
- [4] J. Gao and S. Ye. ACMFS: an abnormal circumstance-driven self-maintenance mechanism based on flexible scheduling. *Proceedings of the International Conference on 3rd Information Sciences and Interaction Sciences (ICIS)*, 318-323, 2010.
- [5] M. Pěchouček and V. Mařík. Industrial deployment of multi-agent technologies: review and selected case studies. *Auton Agent Multi-Agent Syst* (2008) 17:397-431, 2008.
- [6] L. Paulo, V. Paul, and A. Emmanuel. Self-adaptation for robustness and cooperation in holonic multi-agent systems. In Hameurlain A, et al. (Eds.): *Trans. on Large-Scale Data- & Knowl.-Cent. Syst. I*, LNCS 5740, 267-288, 2009.
- [7] E. Di Nitto, C. Ghezzi, A. Metzger, et al. A journey to highly dynamic, self-adaptive service-based applications. *Automated Software Engineering*, 15 (15): 313-341, 2008.
- [8] T. Liu, L. Zhang, B. B. Shi. Adaptive immune response network model. *Emerging Intelligent Computing Technology and Applications: With Aspects of Artificial Intelligence*, 890-898, 2009.
- [9] A. K. Qin, V. L. Huang, P. N. Suganthan. Differential evolution algorithm with strategy adaptation for global numerical optimization. *IEEE Transactions on Evolutionary Computation*, 13 (2): 398-417, 2009.
- [10] R. Charrier, C. Bourjot, and F. Charpillat. Study of self-adaptation mechanisms in a swarm of logistic agents; *Third IEEE International Conference on Self-Adaptive and Self-Organizing Systems, SASO '09*, 82-91, 2009.
- [11] D. Weyns and M. Georgeff. Self-adaptation using multiagent systems. *IEEE Software*, 27(1): 86-91, 2010.
- [12] D. Weyns and T. Holvoet. An architectural strategy for self-adapting systems. *International Workshop on Software Engineering for Adaptive and Self-Managing Systems, ICSE Workshops SEAMS '07*, 3-12, 2007.
- [13] T. Huaglory and U. Rainer. Towards autonomic computing systems. *Engineering Applications of Artificial Intelligence*, 17 (7): 689-699, 2004.

Collective Service Intelligence Management in Mobiquitous Systems

Evgeniya Ishkina

Information Systems department
Astrakhan State University
Astrakhan, Russia
e-mail: ishkina@aspu.ru

Abstract—This paper describes a Self-adaptive Middleware for Mobiquitous information system and its underlying model based upon a multidimensional service representation and management. The service dimensions refer to the “4W”: 1) Who?, i.e., user profile (implicit knowledge deduced from user interaction history and explicit knowledge in form of preferences); 2) Where and 3) When?, i.e., “external” (physical) interaction context (location, time, device, etc.); 4) What?, i.e., “internal” interaction context (user ad-hoc task: goal(s), expectations and optional requirements). An original approach for dynamic adaptive service generation (on-the-fly composition) based on collective service intelligence captured in what (we call “collective services memory”) is proposed. Currently a prototype of the middleware is being implemented for touristic cultural paths in the city of Astrakhan (Russia).

Keywords - service computing; mobiquitous services; middleware; adaptive services; service composition; service mining; context-awareness; NFC standard.

I. INTRODUCTION

We are now entering a new technological era where physical spaces are becoming “smart” and objects – “tagged” thus providing all kind of (*mobiquitous*) services for mobile smart phone holders. *Mobiquity* is a recent word bearing the strategic convergence of *mobility* (mobile phones becoming smart) and *ubiquity* (of Internet becoming local, 2.0 and broadband) [1].

The story of computer science can be seen as a story of functional system/service layers being introduced to ease either application development (successively operating system, database server, application server, mobile server, EDGE Server) or user-friendly convenience. Our proposal consists in creating of a new functional layer on top of the EDGE server that could increase interaction efficiency of users in mobiquitous information systems by proposing them complex services fitting their goals: the interaction system realized in a middleware.

By services we understand web services and NFC (Near Field Communication) mobile services. Services could be different regarding:

- Standards: SOAP (Simple Object Access Protocol) versus REST (Representational State Transfer) standards for web services implementation;
- Interactions: web services with one simple invocation method versus NFC services with its touching paradigm and three operating modes of NFC standard: reader/writer, peer-to-peer, and card emulation mode).

SOAP and REST are two main approaches for building web services. SOAP services are operation-oriented, and RESTful services are resource-oriented. Semantic web services, in addition to simple web services, provide machine-readable semantic data. Existing semantic web service ontologies, such as OWL-S and WSMO, commonly consider only SOAP web services. However, there are proposals for extending them in order to support RESTful web services which are very widespread currently [2]. For NFC services there is not common standard yet.

These heterogeneous services should be consolidated into one integrated warehouse of complex mobiquitous services.

Adaptive information systems are becoming increasingly important especially in the field of mobiquitous information systems. In [3], Sousa *et al.* show that today users are surrounded by technology that is heterogeneous (wide variety of computing platforms, interfaces, networks and services), pervasive (wireless and wired connectivity that pervades most of our working and living environments) and variable (users can move from resource-rich environments, such as workstations, to resource-poor environments, such as a PDA in a park). Mobiquitous systems should be able to adapt to users mobility and system access ubiquity in order to reduce users overhead; they should abstract users from details of access to heterogeneous, pervasive and variable services for maintaining high-level user activities. Users no longer want to get pieces of data, information or knowledge, but they want to get the complete services [4].

Today we can make objects *smart* by tagging them (QR codes, NFC tags), but they still remain reactive only. A mobiquitous system is not adaptive by default; it provides the user with the first level of “intelligence” – the ability to communicate with objects using a mobile phone at *any time*, *anywhere* and on *any device* for obtaining different services. A new functional layer is proposed in our research to make

the ubiquitous system proactive, context (situation) aware, and, thus, to provide users with the second level of intelligence – the ability to obtain an *appropriate customized service*, adapted to the current user *situation* in a transparent manner.

This new functional layer is being implemented within a middleware platform which will hide details of access to heterogeneous, pervasive and variable services from users. Since interaction context is very important in ubiquitous systems, the key idea of this middleware is to manage multiple context-aware strategies (services) for fitting the same user goal, captured by the services usage analysis and then, able to be applied in an appropriate situation. These services can also be recombined in order to construct new useful services with the possibility to evaluate their quality and to infer their functional and nonfunctional characteristics. Each situation of users interactions with the ubiquitous system is unique; therefore, there should be a specific service generated especially for this situation from services fragments and knowledge about their usage stored in the collective services memory.

The remainder of this paper is organized as follows. In Section II, we describe some related research directions; then in Section III, we provide a motivating example of an illustrative scenario. In Section IV, we examine the concept of ubiquitous systems intelligence. In Section V, we present the architecture of the self-adaptive service-oriented middleware and the approach for dynamic service composition. And, finally, in Section VI, we give some research directions to enhance and validate our proposal.

II. RELATED WORK

A. Service composition and service mining

A service is an autonomous IT-asset which can be reused by an arbitrary number of consumers in contexts often unknown at design time. As a matter of fact, services must be designed to be as independent as possible from the context in which they will be utilized [5].

Service composition is an aggregation of multiple services into a single composite service providing more sophisticated functionality and creating add-on values. A user could be provided by a constructor for interactively building his own composite service but a more interesting problem is to automatically recognize user's situation for providing him an appropriate composite service.

In [6], Zheng defines two approaches for service composition. It can be:

- top-down – composition of existing web services driven by specific search criteria, or
- bottom-up – discovery of interesting and useful compositions of existing web services with no such criteria.

And the result of service composition can be:

- one-shot – it realizes a particular request of a particular end-user, or
- reusable – it realizes a generic request of a typical end-user.

In the proposed middleware the bottom-up approach consists in service mining correlated with services dependencies mining, and the top-down approach consists in analyzing goals conjunction, matching corresponding services and their composition. Final services generated by the middleware are always seen as one-shot, but the composition history is saved for further producing a reusable complex service from a set of correlated one-shot compositions.

In [7], Sousa *et al.* present two approaches for services discovery for composition, which can be:

- context-aware – given the context parameters, the suitable services are selected from the service repository, or
- goal-driven – the user makes a request and the system tries to find the most suitable service, which agrees with the request description.

The proposed middleware adopts both approaches: first, the services are selected by their goals annotations, and then the engine looks for the most suitable service composition taking into account context parameters.

Our aim is to deal with composability of different kind of heterogeneous services (mobile NFC services, SOAP/RESTful web services, semantic web services, etc.) through a middleware by integrating them at structural and semantical levels and by mining collective service intelligence.

B. Context-, situation- and task-awareness

Context awareness is referred to the capability of an application or a system to be aware of its physical environment and situation in order to be able to act and answer in a proactive and intelligent way [8].

In the field of services, context was often seen only as user location attached to popular location-based services. In our proposal we consider a larger context concept which is *situation* – combination of system-side, user-side and environment-side parameters.

In [3], the concept of task-awareness is presented. It means carrying out high-level users activities: planning a trip, buying a car, etc. In today's systems those activities and goals are implicit. In task-aware systems, users specify their tasks and goals, and it is the responsibility of the system to automatically map them into the capabilities available in the ubiquitous environment.

In our proposal the task corresponds, from one side, to a combination of goals, and from another side, to a set of service compositions. The system learns from usage how to make a reasoning on combined goals and learns about the strategies of fulfilling the goals, i.e., service scenarii.

C. Overview of existing approaches for intelligent service composition and delivery

Let consider three main groups of approaches for intelligent service composition and compare them with the proposal.

The first group is represented by approaches for context-aware service composition [7][9].

iCas [7] is a service-oriented architecture that uses an open ontological context model (SeCoM) to provide personal and contextual information and to support the composition of context-aware services on the fly. A prototype of the iCas platform is implemented. When starting services composition, user can add and remove services interactively. Possibilities for composition are returned to user based on the current context and user policies. Services are described using OWL-S.

MyCampus [9] is a semantic web environment aimed at enhancing everyday campus life. Users acquire or subscribe to a variety of task-specific agents that assist them in the context of different tasks. MyCampus supports the dynamic discovery and access of contextual information sources and the automated generation of plans by task-specific agents through the discovery of services that can be dynamically composed to satisfy one or more user-goals.

There also exist some approaches that do not allow service composition but they aim at delivering of context-aware services. For example, SOCAM [10] is a middleware architecture that supports the building and rapid prototyping of context-aware services. It uses the formal context model based on OWL to represent, manipulate and access context information.

Another group of approaches lie in the field of goal-driven service composition [11][12].

In [11], a goal-driven approach for service composition is presented. The authors propose a task-oriented semantic representation model of web services and based on this model goal-driven service composition is performed dynamically to achieve user's goal. The relevant concrete web services to complete the task are bound dynamically in the runtime.

In [12], a goal-based approach for dynamic service discovery and composition is described. The approach is based on a behavior model represented by goal modeling. Goals can be further decomposed into sub-goals, and tasks fulfill (sub-)goals. This approach is founded in a well-defined set of domain and task ontologies.

The third part of approaches concerns pattern-driven service composition. In [13], Tut *et al.* propose the use of patterns combined with the domain knowledge for facilitating the composition process of e-services. Patterns represent a proven way of doing something, "a three-part rule, which expresses a relation between a certain context, a problem and a solution". An example of instantiating of generic patterns into specific ones is presented.

Approaches for context-aware and goal-driven composition are often well separated, however there exist some mixed approaches, for example [9]. The majority of approaches for pattern-driven composition are not context-aware. Patterns represent result of an attempt to find context-free service sequences.

Our proposal consider pattern as a service: it has the same characteristics, it can be annotated by situation elements, it can make part of service composition. Patterns reveal relationships between services, and along with the context (situation) dependences represent collective service intelligence.

To our knowledge, there are not mixed approaches in the field of web services and interactive NFC services. Ubiquitous services which are being considered in this paper integrate smart objects with related web services and are consumed in a high interactive manner. However there are some research results in the field of context-aware NFC applications [14].

III. USE CASE SCENARIO

For better understanding of the remainder of this paper we will provide a use case scenario.

Consider two ubiquitous services in the city of Nice, in France (an NFC European city since May 2010). The first one is a complex NFC mobile service offering a guided tour on the "invisible historic cultural path" of Gogol, Russian writer, in Nice consisting of all places where he lived in the 1840's associated with his name and providing multimedia information attached to tags (QR Code and NFC tags) with the possibility to produce information on each point of the path and interface with social networks [15]. The second one is a web service of restaurant booking.

The complex goal of Ivan, a Russian tourist, is to make a complete tour and to have lunch in a restaurant (probably by making a break in his tour). His goal remains persistent during an interaction session and corresponds to initial interaction constraints; if Ivan would like to change his goals, a new session will start.

For the given goal we should consider two elements of the external environment: location and time. These are not constant during the interaction, their values are generated in real time.

Concerning the user himself, there are some extra parameters which could influence the system behavior. These are Ivan's meal preferences. If Ivan did not provide them to the system, the system itself can infer them based upon analysis of Ivan's interaction history. If there are not enough data for such analysis, the system can rely on the nationality of Ivan (if known) and use this information in conjunction with learned dependencies between nationalities and meal preferences. Let us consider that in this example meal preferences of Ivan are not known but the system knows that he is Russian and that most of Russian tourists prefer Russian restaurants in Nice.

Now we will describe some possible options. For example, one of the point of Gogol path concerns his preferred restaurant of French cooking. And the system has learned that whatever specific user preferences are, if the user goal is to have a lunch within this guided tour, they generally select the restaurant that Gogol preferred and not the one related with their own preferences.

So, Ivan starts his tour "Gogol in Nice"; he loads the special application or information on his mobile phone by touching a given touristic NFC tagged poster. The system proposes him to complete his goal with some specific goals, one of them is having a lunch – and he selects it. In the middle of his tour Ivan decides to have a lunch break. The system then generates the list of recommended restaurants. If the Gogol preferred restaurant is nearby to the Ivan's location, it will be the most recommended. If not – the

system selection will be based upon Ivan's meal preferences primarily. If Ivan doesn't make a break but he is already in the Gogol preferred restaurant, the system provides him a proposition without his explicit demand.

IV. MOBIQUITOUS SYSTEMS INTELLIGENCE

In ubiquitous systems, there exists an augmented need of tailored application delivery for reducing user overhead in ubiquitous environment that implies the necessity of discovering implicit collective intelligence of ubiquitous services and providing multidimensional usage view.

A. Collective service intelligence

In [16], O'Reilly *et al.* present collective intelligence concept evolution. They introduce the concept of *Squared Web* which play an intermediate role between Web 2.0 (social) and Web 3.0 (semantic). Web 2.0 offers to users the possibility to generate the content. The future Web 3.0 will allow to machines the possibility to understand data but it requires a lot of work for the current Web semanticization. Web 2.0 focuses on collective human intelligence, while Squared Web focuses on collective intelligence of captors, tags, etc.

In ubiquitous systems we can consider collective intelligence of services. Services cannot be fully structured at design-time because there are still many unknown dependencies of users and usage contexts. The additional functional layer we are proposing, should manage the collective service intelligence and learn automatically about services usage in particular situations for further better interaction efficiency.

B. Multidimensional usage representation

Mobiquitous services usage can be represented in multidimensional space. The two typical (minimum) usage dimensions are user profile and location. Thus, users can be provided with customized services taking into account these parameters. But each service provider can make a reasoning on it, in its own way.

The idea to tag real objects in space is not new. It has been widely used in tracking objects in logistics, etc. Users interacts with a ubiquitous system via tagged and location-based objects; the system can give some recommendations to users based upon where he is located (and propose him restaurants, museums, shops, etc.). Here, we have two levels of context: the first one corresponds to the information which is directly obtained from sensors (spatial coordinates) and the second one corresponds to the inferred information (available services in users' vicinity).

The relatively new idea is to tag real objects in time scale. Saving and analyzing interaction tracks allow to get information about when, how and by whom these objects were used.

Rules applied for generating recommendations can be static (anyone who touches an NFC poster at the bus stop receives the same list of restaurants in the neighborhood), or customizable (based upon explicit user meal preferences or implicit preferences inferred by the system from the user interaction history).

In the area of ubiquitous systems we consider three usage dimensions (dimensions of context in its global meaning which we call *situation*):

- Interaction actor – user profile (explicit and implicit knowledge), we will call it user dimension;
- “External” (physical) interaction context (location, time, etc.), we will call it context dimension;
- “Internal” interaction context (user task: goal(s) and constraints), we will call it task dimension.

Depending on the application area (m-tourism, m-marketing, etc.) the importance of different dimensions varies, but there exists one common feature – ubiquitous systems are *task-driven*. The user looks for a complex service that fulfills his goal while hiding most part of implementation details; he does not look for isolated data, information, knowledge and services fragments. User tasks and goals are hierarchical and multiple service scenarios for achieving them are discovered at run-time. If there is none on-the-shelf solution in the collective services memory, the user goals are then decomposed in order to find matching services for sub-goals. The remaining two situation dimensions (user profile, physical environment) are used to select the most appropriate scenario for a given situation.

Thus, ubiquitous systems intelligence is based upon answering the following questions:

- When, where, how and by whom ubiquitous services should be used? – *Learning about usage situations.*
- How services can interact with one another? What service compositions could be useful to users and what is the typical situation profile for this? – *Performing service mining with corresponding situation parameters mining impacts appropriate services selection.*
- How to evaluate situations equivalence in a flexible manner depending upon application areas? – *Analyzing services used in these situations with some common parameters.*

V. SELF-ADAPTIVE MIDDLEWARE FOR MOBIQUITOUS SYSTEMS

A. Logical architecture

In Figure 1, the global centralized logical architecture of the self-adaptive middleware for ubiquitous systems is presented. Below we describe its major components.

1) Atomic services integration.

This layer enables integration of heterogeneous ubiquitous back-end services (SOAP/RESTful web services, semantic web services, NFC mobile services) by providing a unique service metamodel. It is a service access layer: only at this level details of heterogeneous services invocations are known.

2) Collective services memory management.

This layer enables discovery of new services compositions useful in particular situations, and enables management of all services – atomic and composite.

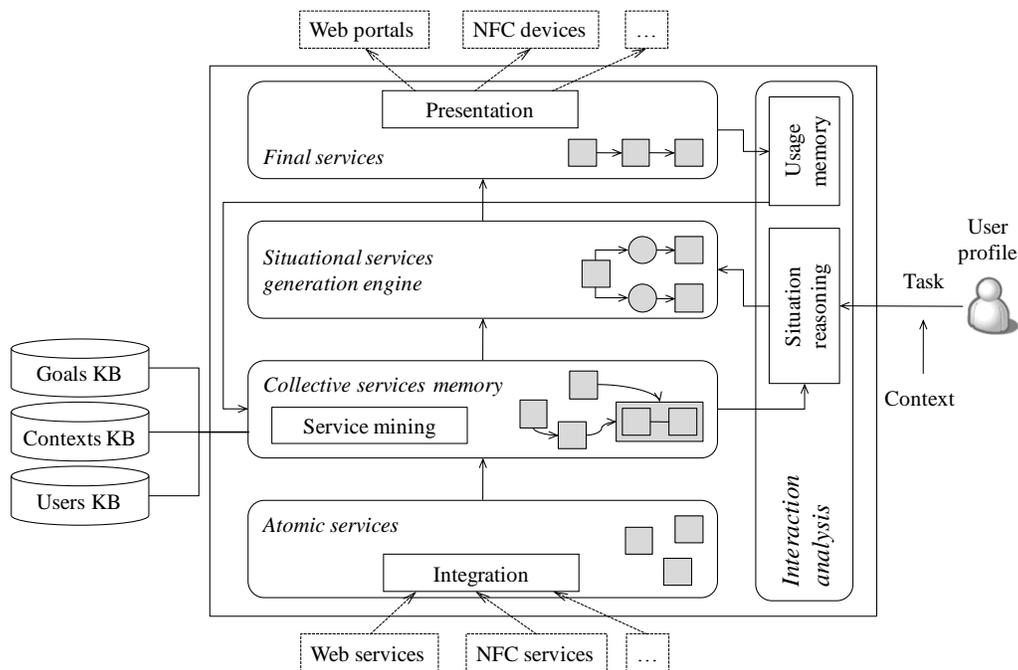


Figure 1. Logical architecture of self-adaptive middleware for ubiquitous systems

At this level the second layer of services metadata is defined, it represents situation parameters (task, user, context) learned from interaction history by applying data mining techniques.

“Collective” means that all services are stored in the memory along with explicit and implicit relationships between them. Explicit relationship between two services is their joint use in a composite service. This use is conditional, related to the appropriate usage situation. Implicit relationship between two services consists in similarity of situations of their appropriate usage. Both explicit and implicit relationships between services are processed for pre-selecting services for final service generation.

Basic elements of collective memory are services (atomic or composite) which represent usage fragments. These fragments are used in the next layer for generating a unique service corresponding to user’s situation.

All atomic services with their metadata generated at the integration level are stored within the collective memory. Then, it is enriched by compositions of existing services obtained by service mining algorithms.

Special algorithms allow evaluation of the usefulness of discovered service composition, i.e., the probability of its reuse, it also allows predicting of functional and non-functional characteristics of composite services.

Special algorithms are also used to mine key situation parameters which may influence the use of services together in a scenario. These parameters represent preconditions taken into account during adaptive end-user service generation.

Each of the three dimensions of usage situations (task, user, context) correspond to an ontology constructed during the system functioning for reflecting the set of situational

parameters and their importance for a given application domain.

Service itself is independent from usage situations, it has a goal to achieve. Multidimensional service annotations based on task, user, and context ontologies enable evaluation of service relevance for the given situation.

Thus, collective services memory stores usage fragments – atomic services and useful composite services both annotated by ontologies corresponding to situation dimensions. Atomic services at this layer are considered as abstract elements described using a unique service metamodel without any details of their invocation.

3) Situational service generation.

Taking into account the situation description and services metadata stored in the collective memory, this layer generates a unique service corresponding to the given situation and representing a composition of services fragments from collective memory.

Some details of our approach for situational service generation are given in the next subsection.

4) Final services presentation

The purpose of this layer is to interact with atomic services integration layer for invocation of composition elements. This layer also provides the user with an appropriate interface.

5) Interaction analysis

Usage history component allows saving system usage logs and their preprocessing for further use of service mining algorithms.

Situation reasoning component hides from all other middleware components details of capturing 1st level situation data, i.e., information which is directly obtained from sensors. It constructs the 2nd level of situation

representation, i.e., high-level information inferred from the 1st level situation data and related with task, user and context service dimensions operated by the middleware. It is further processed by the situational service generation engine.

B. Approach for situational service generation

Our approach for dynamic service generation is based upon the use of ontologies (Figure 2).

First of all, there is a domain ontology for sharing vocabulary between all middleware components. Then, there are three ontologies corresponding to *users*, *contexts*, and *goals* using the domain ontology. Goal ontology is totally domain-specific; for user and context ontologies some domain-independent elements can be defined.

Services in the collective memory are annotated by all the three dimensions ontologies. The situation is composed with task, user, and context descriptions. Task represents a set of goals indicated by the user. All useful parameters of user profile and physical context are preprocessed at the interaction analysis layer thus constructing the 2nd level of situation. User's and context's descriptions are based on the corresponded ontologies.

Situational service generation engine receives situation description and annotated services from collective memory.

Service discovery for composition is task-driven, it is based on task goals matching. In case of no services matched, task goals are decomposed for trying to find services corresponding to sub-goals.

Candidates are tested for correspondence to the given situation and for syntactic and semantic compatibility. The final service is then composed of the most appropriate services fragments.

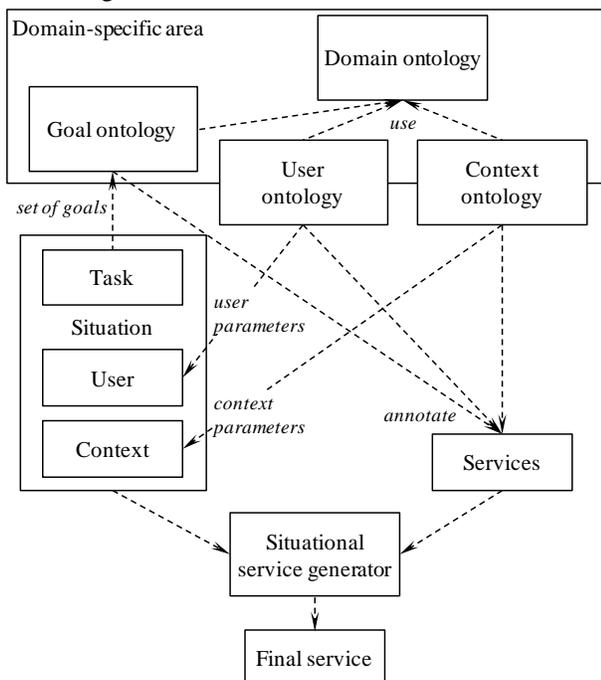


Figure 2. Situational service generation

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented the self-adaptive middleware for ubiquitous systems which manages the intelligence of ubiquitous information systems along with a description of its major components. We also described our approach for situational service generation, i.e., adaptive composition. The basic rationale is that, in distributed heterogeneous systems, it is necessary to collect knowledge about usage. This knowledge should not be the simple facts, but it should represent strategies of goals achievement in particular situations. Users no longer want to get isolated information; they look for complex services. Current systems are becoming task-aware: their goal is to match high-level user task in a transparent manner. Services could then be recombined for producing new ones with the possibility to evaluate the important characteristics of service composition, its usefulness in particular situations.

One of current tasks consists in analysis of the middleware architecture using available software engineering methods such as ATAM (Architecture Tradeoff Analysis Method).

The proposed approach is being implemented and validated for touristic cultural paths in the City of Astrakhan (Russia). Finally, our proposed architecture described here, will be formally described in companion research papers.

ACKNOWLEDGMENT

The author would like to express her appreciation to two Professors of Computer Science from the University of Nice Sophia Antipolis (France) and I3S Research laboratory, Serge Miranda for his valuable comments and remarks during this paper preparation, and Nhan Le Thanh, for his research guidance and impetus.

REFERENCES

- [1] S. Miranda and E. Ishkina, "NFC ubiquitous information service prototyping: overview, lessons, state of the art, innovation and research directions", 3rd International Conference on Information Systems and Economic Intelligence (SIEE'2010), Sousse, Tunisia 18-20 February 2010.
- [2] O.F. Ferreira Filho and M.A. Grigas Varella Ferreira, "Semantic Web Services: a Restful Approach", IADIS International Conference WWW/Internet, Rome, Italy, 2009.
- [3] J.P. Sousa, V. Poladian, D. Garlan, B. Schmerl, and M. Shaw, "Task-based Adaptation for Ubiquitous Computing", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Special Issue on Engineering Autonomic Systems, 36(3), 2006, pp. 328-340.
- [4] G. Zheng and A. Bouguettaya, "A Web Service Mining Framework", Proc. of the Int. Conf. on Web Services (ICWS'07), Salt Lake City, Utah, USA, 2007.
- [5] J.J. Dubray, WS-PER: An abstract SOA framework <http://www.wsper.org/> (2007)
- [6] G. Zheng, "Web Service Mining", PhD Thesis, Virginia Tech, 2009.
- [7] J.P. Sousa, E. Carrapatoso, B. Fonseca, M.G.C. Pimentel, and R.B. Neto, Composition of context aware mobile services using a semantic context model, International Journal on Advances in Software, 2:2-3, 2009, pp. 275-287.
- [8] A.K. Dey and G.D. Abowd, "Towards a better understanding of context and context-awareness", Proc. of the Workshop on the What,

- Who, Where, When and How of Context-Awareness, ACM Press, New York, 2000.
- [9] M. Sheshagir, N. Sade, and F. Gandon, "Using Semantic Web Services for Context-Aware Mobile Applications", in *MobiSys 2004 Workshop on Context Awareness*, Boston, 2004.
- [10] T. Gu, H. Pung, and D. Zhang, "A service-oriented middleware for building context-aware services", *Journal of Network and Computer Applications*, 28(1), 2005, pp. 1-18.
- [11] K. Zhang, Q. Li, and Q. Sui, "A goal-driven approach of service composition for pervasive computing", in *Proc. of the 1st International Symposium on Pervasive Computing and Applications*, 2006, pp. 593-598.
- [12] L.B. da Silva Santos, L.F. Pires, M. van Sinderen, "A Goal-Based Framework for Dynamic Service Discovery and Composition", *International Workshop on Architectures, Concepts and Technologies for Service Oriented Computing*, Porto, Portugal, July, 2008, pp. 67-78.
- [13] M.T. Tut and D. Edmond, "The Use of Patterns in Service Composition", *Proceedings of the International Workshop on Web Services, E-Business, and the Semantic Web*, Toronto, Canada, 2002, pp. 28-40.
- [14] P.C. Garrido, G.M. Miraz, I.L. Ruiz, and M.A. Gómez-Nieto, "A Model for the Development of NFC Context-Awareness Applications on Internet of Things", *2nd International Workshop on Near Field Communication*, Monaco, 2010.
- [15] E. Ishkina and S. Miranda, "NFC Ubiquitous Ecosystem for Information Services of the Future: Applications to M-tourism and M-learning", *Proc. of International scientific conference "Electronic culture. Information technologies of the future and modern e-Learning"*, Astrakhan, 6-8 October 2009.
- [16] T. O'Reilly and J. Battelle, "Web Squared: Web 2.0 Five Years On", *Web2.0 Summit Conference*, O'Reilly Media, San Francisco, 2009, pp. 1-13.

Service Network Modeling and Performance Analysis

Manolis Voskakis
 Transformation Services Lab
 University of Crete
 Crete, Greece
 e-mail: voskakis@tsl.gr

Christos Nikolaou
 Transformation Services Lab
 University of Crete
 Crete, Greece
 e-mail: : nikolau@tsl.gr

Willem-Jan van den Heuvel
 European Research Institute in Service Science
 Tilburg University
 The Netherlands
 e-mail: wjheuvel@uvt.nl

Marina Bitsaki
 Transformation Services Lab
 University of Crete
 Crete, Greece
 e-mail: bitsaki@tsl.gr

Abstract — Services play an important role in business interactions among partnerships forming value-creating service networks. A central problem in service network design is to analyze participants' behavior and optimize their value. In this paper, we propose a simulation model to evaluate the long-term impact of changes to resources and predict the performance of service networks. Successful predictions of the future behavior of service networks help analysts improve service network's functionality.

Keywords- service networks; value optimization; performance analysis

I. INTRODUCTION

The growth of service economies coupled with the evolution of information technology have increased the complexity of service companies in a world of interactions and partnerships. We observe that large and vertically integrated firms are replaced by value-creating service networks. Service networks consist of interdependent companies that use social and technical resources and cooperate with each other to create value [1], [2], [3].

Fig. 1 depicts the anatomy of a car repair service network comprised of five interrelated levels. In particular, the top level defines end-to-end processes connecting service provisions of several service providers (Original Equipment Manufacturer (OEM), Car Dealers and Clients [4]). In this way a service network can be partitioned into a set of discrete business services that completely process service client requests. Fig. 1, shows that an end-to-end process such as car repair is subdivided into composite service processes such as diagnosing the problem to be repaired, ordering part replacements and perform the repair. The order process shown in Fig. 1 is a composition of several atomic services (see corresponding level) such as investigating failure symptoms, identifying parts, ask advise from technicians, and ordering the appropriate (possibly upgraded) parts. Software and human services can be

routinely mapped to atomic services, and can be selected, customized and combined into aggregated service applications. The software service may be deployed on a software service infrastructure, which may for example be a distributed cloud environment, providing the capabilities required for enabling the development, delivery, maintenance and provisioning of services as well as capabilities that monitor, manage, and maintain QoS such as security, performance, and availability.

Clearly, the trend will be to move to high-value service networks where business process interactions give rise to new service analytics models and techniques that will help to pro-actively manage services and pinpoint areas for improvement.

Various approaches have been proposed to measure the performance of service networks [5], [6], [7]. Most of the research has focused on describing models that represent inter-organization exchanges. In [5], a quantifiable approach of value calculation is proposed that connects value with expected revenues. In contrast, Biem and Caswell [6] describe building block elements of a value network model and design a network-based strategy for a prescriptive analysis of the value network. Allee [7] provides a systematic way for approaching the dynamics of intangible value realization, interconvertability, and creation. Biem and Caswell [6] and Allee [7] use qualitative methods to describe value in a service network in contrast to Caswell et al. [5] that calculates that calculates value in a quantifiable manner. The above approaches do not study strategic behavior of network participants that would result in value optimization.

In this paper, we study the impact of strategic changes on the performance both at the level of the network as well as its participants. In particular, we introduce an analytical model and associated simulation tool to optimize value. Comparing to previous work that has been done, we improved the estimation techniques and we used a powerful

simulation tool to perform our experiments and analyze dynamic “what-if” questions such as: what is

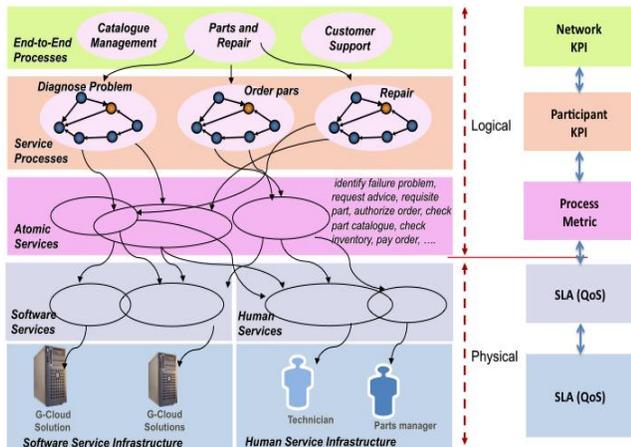


Figure 1. The anatomy of service networks.

the impact of setting optimal – for one participant - prices on the performance of the other participants as well as the entire network? What is the impact on the performance if a new participant suddenly enters the service network? Are there any equilibrium strategies among the participants that eliminate their conflicts of interests?

We extend the model presented in [5] and take into consideration the expected costs to estimate the expected value of the network and the various participants. We also improve the methodology used in [5] to provide estimations of revenues and satisfaction measures. Our main contribution is the definition and solution of value optimization problems with respect to service prices.

We observe that participants’ value depends on their expected profits. Expected profits express the additional value that will be accrued by the relationship levels a participant develops when it sells goods and services to other participants or to the end customers. This value is related to its intangible assets and on the degree of satisfaction it obtains from its customers. There are many approaches that have been proposed to measure customer satisfaction. In this paper, we use the methodology proposed by Fornell et al., known as American Customer Satisfaction Index [8].

We use the System Dynamics approach [9], [10] to analyze the behavior of a complex system (car repair service network) over time. System dynamics tools allow modelers to succinctly depict complex (service) networks, visualizing processes as behavior-over-time graphs, stock/flow maps, and causal loop diagrams. These models can be tested and explored with computer simulation providing for example better understanding of the impact of policy changes (e.g., through animation of (service) systems) and facilities for sensitivity analysis. Examples of such tools include iThink [10], Vensim [11] and PowerSim [12].

In this paper, we have adopted the iThink tool to investigate the fluctuation of value under different circumstances. The results of these simulations provide predictions about the future of the service network in order to increase its adaptability to the changes of the environment and enable network participants to determine the most profitable co-operations and attract new ones. We show that the interactions among the participants of a network force them to reach equilibrium otherwise the network will collapse.

The remainder of this paper is organized as follows: Section II describes the car repair service system. Section III presents the methodology proposed to estimate value in service systems. In Section IV, we analyze the case study and run experiments to measure its performance. The results of the simulations are presented in Section V. Finally, in Section VI, we provide some concluding remarks.

II. MOTIVATING SCENARIO

The motivating scenario revolves around a service network that links four types of participants: an Original Equipment Manufacturer (e.g., Volvo), Car Dealers (with repair facilities), Suppliers and Customers. In particular, the scenario considers the end-to-end process “Order & Repair” that was already briefly introduced in the introduction.

The scenario that we will use during the remainder of this article is an extension to [5] and basically looks as follows. OEM-franchised dealers may service and repair cars for their clients. Both activities require a car parts catalogue to ensure that repairs can be performed efficiently either in the replacement of parts or repairing after accidents. The part catalogue facilitates efficient installation, operation and lifecycle maintenance of intricate products describing detailed part information that can be fully integrated with other service applications supporting customer support processes, human resource management, and other service provisions.

The quality of the OEM parts, catalogues, and OEM support services influences how many OEM parts will be ordered and used for a car repair and how many parts will be used from Third Party Suppliers (TPS), and how many customers will go to OEM dealers or to TPS dealers. OEM obtains parts from certified supply-chain suppliers (SCS).

The technicians report the car service requirements that may include replacing teardowns, warranty replacements and collision repairs. On the basis of the car diagnosis, a cost estimate will be computed and communicated to the client for authorization. Once authorized the automotive technician will scrutinize failure symptoms, detect faulty parts, order parts and perform the repair. Ordering parts is a complex process that involves asking advice from expert technicians from the OEM, including acquiring information about parts under warranty, and getting approval from the dealer’s part manager. The part manager then checks local inventory for the required part, and if necessary checks the stock at the OEM or supplier stocks, and eventually places an order. The

part manager may either use third-party suppliers or suppliers from certified supply-chain suppliers.

III. THE MODEL

In this section, we introduce our service performance analytics model in support of strategic analysis of service network changes and improvements. Theorizing on service networks, and particularly performance analysis, can be addressed from multiple and often complementary perspectives. In our work, we propose a methodology to calculate value in service systems. We focus on the dynamic environment in which service networks emerge, and especially on connectivity and profitable cooperation that play an important role in value creation. We use our model to investigate network profitability and give answers to the following:

- Determine the conditions under which it is profitable for a firm to participate in the network and identify the factors that influence its value.
- Identify key stone participants (participants that create the most value for the network).
- Determine participants' optimal strategic decisions (cooperating with someone or not, joining the network or not, etc.).

We consider the service network as a set B of participants connected through transfer of offerings that delivers value to them. All offerings are treated as services that are composed by participants' interactions and co-operations to provide a final service to a set C of end customers. Let p_{ij} denote the price participant i charges participant j for offering its services and r_{ij} denote the service time of the interaction between participants i and j . Price and time are the main parameters that affect customer satisfaction which is in turn the corner-stone for calculating value as we will see below.

A. Customer Satisfaction

Customer satisfaction measures the willingness of end customers to buy the services offered by the network and influences the increase or decrease of new entries. The calculation of satisfaction $SAT_{ij}(T_N)$ of participant j for consuming services from participant i at the end of the time interval $[T_{N-1}, T_N]$ for our model is a variation of the American Customer Satisfaction Index (ACSI) [9] and basically described as follows. ACSI is operationalized through three measures: q_1 is an overall rating of satisfaction, q_2 is the degree to which performance falls short of or exceeds expectations, and q_3 is a rating of performance relative to the customer's ideal good or service in the category. Without loss of generality, we quantify the above measures using the following formula:

$$q_k = [(\beta_k/p_{ij})0.6 + (\gamma_k/r_{ij})0.4], k=1,2,3, \quad (1)$$

where $[x]$ denotes the integer part of x and β_k, γ_k are the parameters that determine the effect of price p_{ij} and time t_{ij} respectively on q_k . In our analysis, we use the following function (see [8] for further details) to calculate the satisfaction:

$$SAT_{ij}(T_N) = (w_1q_1 + w_2q_2 + w_3q_3 - w_1 - w_2 - w_3) / (9w_1 + 9w_2 + 9w_3), \quad (2)$$

where w_k are weights that indicate the importance of each measure q_k .

B. Participants' Value

We consider that an economic entity within a service network has value when it satisfies the entity's needs and its acquisition has positive tradeoff between the benefits and the sacrifices required. We emphasize on the gains or losses captured by the relationships between participants in order to compute value. Our focus is on the methodology in [5], but with a different view of the utilization of relationships between the participants. We define the expected profits $Ep_{ij}(T_N)$ of participant i due to its interaction with participant j to be the expected value of participant i in the next time interval $[T_N, T_{N+1}]$ increased (or decreased) by the percentage change of the expected satisfaction $ESAT_{ij}(T_N)$ in the next time interval and is given by:

$$Ep_{ij}(T_N) = (ESAT_{ij}(T_N) / ESAT_{ij}(T_{N-1})) (ER_{ij}(T_N) - EC_{ij}(T_N)), \quad (3)$$

where $ER_{ij}(T_N)$ and $EC_{ij}(T_N)$ are the expected revenues and costs respectively for the next time interval. Thus, the value $V_i(T_N)$ of participant i at the end of time interval $[T_{N-1}, T_N]$ is the sum of its revenues and the expected profits minus the costs that come from its relationships with all other participants. The total value of the network is the sum of the value of each participant.

C. The Mechanism for Value Calculation

In this subsection we present our value-based model that provides a mechanism to calculate value divided in various hierarchical levels. Fig. 2 (generated by iThink) shows the upper level of the hierarchy and visualizes the basic elements of our framework. We use the example of Section II to simplify our description. Each node represents a module that calculates the value of a participant. Arrows represent dependencies between modules. Each module encloses a sub-system that calculates the value of the module (second hierarchical level). Complex variables inside the module are presented as modules too. Fig. 3 shows the dealer's value calculation process. The green arrows show the impact a module has on another module (e.g., dealer's expected profits increase as dealer's revenues increase). The module dealer's cost in the third hierarchical level is depicted in Fig. 4.

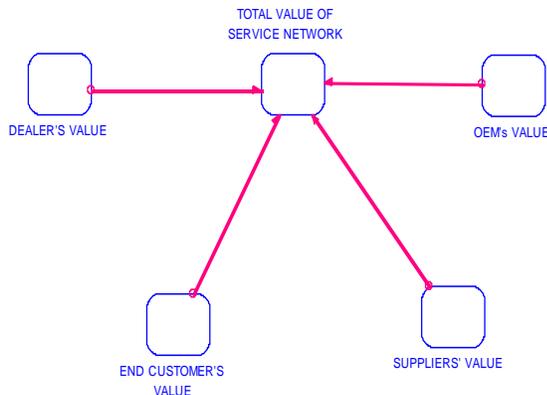


Figure 2. First hierarchical level of value mechanism.

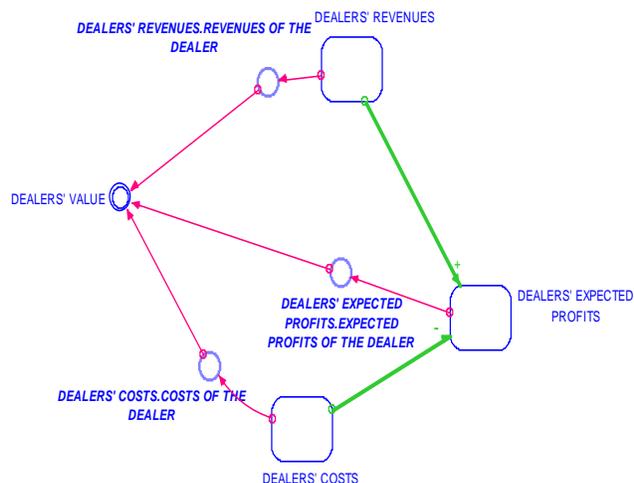


Figure 3. Second hierarchical level – dealer's value.

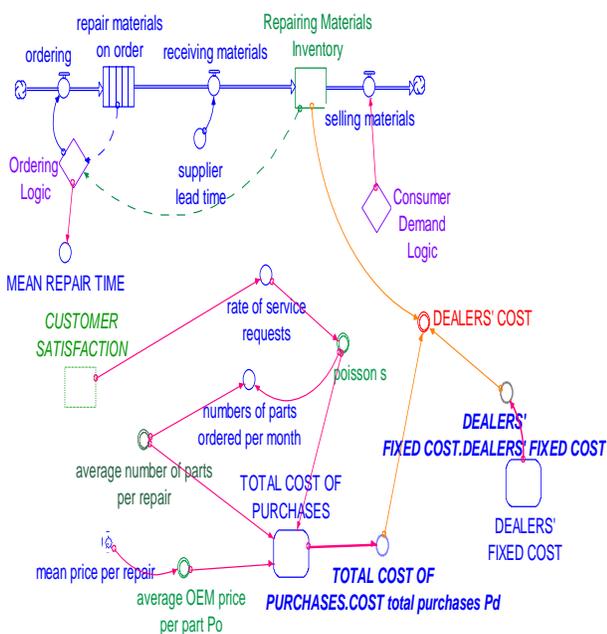


Figure 4. Third hierarchical level – dealer's cost.

IV. SIMULATION EXPERIMENTS TO THE CAR REPAIR SERVICE SYSTEM

We perform simulation experiments to analyze our model making use of 4 scenarios. First, we apply our approach to the car repair service system (Section II) to examine the network's evolution over time. We represent technicians, the parts manager, and the help desk experts as economic entities, each of which is offering its labor as a service to the service system. We measure rates of offerings and payment flows per month over a period of about 30 months. End customer service requests denoted by s are strongly affected by end customer satisfaction, since satisfied customers attract new customers to enter the network. Without loss of generality, we consider that the service requests are produced by the Poisson distribution with mean es being the output of the function:

$$es = -a_1SAT^2 + a_2SAT, \tag{4}$$

where $a_2 > 2a_1 > 0$ so that es is an increasing function of SAT in the range $[0,1]$. (We have chosen (4) because the rate of increase of es decreases with respect to SAT .) We also consider that the number of technicians is a function of the number of service requests; we take that the number of technicians increases linearly with the number of service requests. We calculate the value of each participant as a function of price and time and determine its optimal level with respect to price. The equations of revenues and costs for the dealer, the OEM and the suppliers are taken from [5]. As opposed to [5] that calculates value for a given mean repair price and time, we optimize value with respect to mean repair price and time.

Second, we use the transformation of the basic model, as in [5], in order to cut costs and increase value. Concisely, a solution provider achieves interoperability between participants' information systems through application software operated by the OEM. The application allows everyone to have access to up-to-date information about parts at any time, as soon as this information becomes available to the data base of the application. The gain from the new IT infrastructure is twofold: repair time is reduced resulting in customer satisfaction increase and OEM's mailing costs are eliminated. We apply our methodology to the transformed network to show that the continuous changes of the environment push the network to restructure itself in order to remain competitive. We determine the time interval in which we observe positive effects in profitability in the transformed network compared to the initial one. We also determine which of the participants benefit from the transformation and which not.

Third, we consider a model in which the group of dealers is replaced by a new one that offers more complementarities to the end customers without increasing the mean repair price. This action seems to be profitable due

to the increase of the satisfaction of the end customers of the service network. However new dealers have higher costs that may affect service network’s value. We examine the value of these dealers and the value of the entire service network provided that OEM chooses to cooperate with them.

Fourth, we investigate Nash equilibrium strategies [13], [14] between OEM and the dealer. We define as a strategy for OEM and the dealers the mean profit rates a and b of selling parts and repair services respectively. Let p_s, p_0, p_d be the mean prices set by the suppliers, OEM and dealers respectively for offering their services. Then it holds that:

$$p_0 = p_s + ap_s = (1+a)p_s, \tag{5}$$

$$p_d = p_0 + bp_0 = (1+b)p_0. \tag{6}$$

We examine the existence of equilibrium strategies considering that the rest of the network participants (apart from OEM and the dealer) do not affect their decisions. We assume that OEM buys parts from certified suppliers at a given price p_s .

V. RESULTS

In this section, we present the simulation results from our analysis. First, we compare the basic model with the transformed one.

A. Value Optimization in Basic and Transformed Network

We show the mean repair price p^* that maximizes the dealers’ and OEM’s value in Table I.

TABLE I. COMPARISON BETWEEN THE BASIC AND THE TRANSFORMED NETWORK

Value	Model			
	Basic Network		Transformed Network	
p^*	111 (dealer)	225 (OEM)	116 (dealer)	218 (OEM)
Dealer	51.469.012	34.700.000	46.874.332	34.985.000
OEM	$8500 \cdot 10^6$	$26793 \cdot 10^6$	$9100 \cdot 10^6$	$29990 \cdot 10^6$

We observe that:

- The dealers’ optimal mean repair price in the basic service network is lower than in the transformed service network, since the mean repair time (that affects value) decreases, so the dealer charges his customers less. Consequently, the dealer is forced to increase the mean repair price in order to increase its revenues. Nevertheless, at the optimal mean repair price, dealers’ value is less in the transformed network since the customer satisfaction has decreased as well (higher charges).
- OEM’s value is much higher in the transformed network than in the basic one. This is explained by the fact that the mean repair time decreases and the customers are more satisfied (at OEM’s optimal mean repair price). In addition, OEM in the

transformed network has much lower mailing and labor costs.

- In both networks OEM’s value at dealer’s optimal mean repair price (111 and 116 respectively) is very low compared to OEM’s value at his optimal mean repair price. This means that OEM will never be satisfied to offer its services at prices that reach dealer’s optimal level.
- Dealers’ value at OEM’s optimal mean repair price is higher in the transformed network, since OEM’s optimal price is lower (218).

Furthermore, the simulation results show that, OEM’s value in the transformed network is not higher than that of the basic network from the first month. It dominates after 10-12 months, when both networks offer their final services at their optimal mean repair price (Fig. 5). When both networks offer their services at common prices in the range of 80 to 350, the transformed network dominates the basic network at month 8 to 17.

Finally, the total value of the transformed network (32.190.040.300) is maximized at mean repair price 216 and

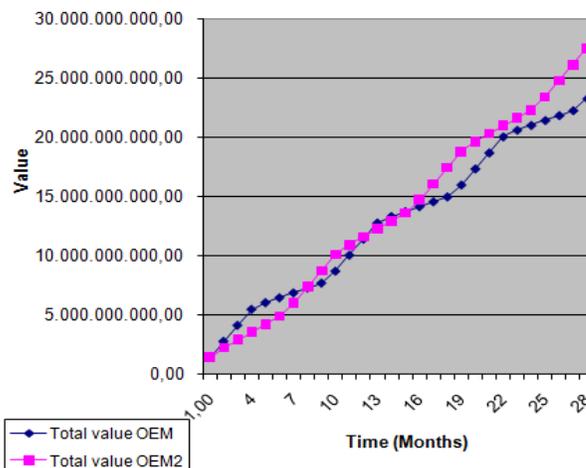


Figure 5. OEM’s value in basic (1) and transformed (2) network at common mean repair prices.

is higher than that of the basic network (28.593.400.000) which is maximized at mean repair price 223. This is explained due to the fact that end customers are more satisfied and OEM (the keystone participant) has managed to cut costs at a great extend in the transformed network. Moreover, we see that the optimal mean repair price for both service networks is very close to the optimal mean repair price of OEM, since OEM contributes the largest part of the total value of the network.

B. Sensitivity Analysis of the Mean Repair Price

In this section, we investigate the impact of mean repair price changes to the dealers’ value. As the mean repair price increases, the difference between the dealers’ value in the basic network and that in the transformed network is

smaller. This is justified by the fact that although the service requests decrease the mean repair price increases resulting in a decrease of the total value as shown in Fig. 6.

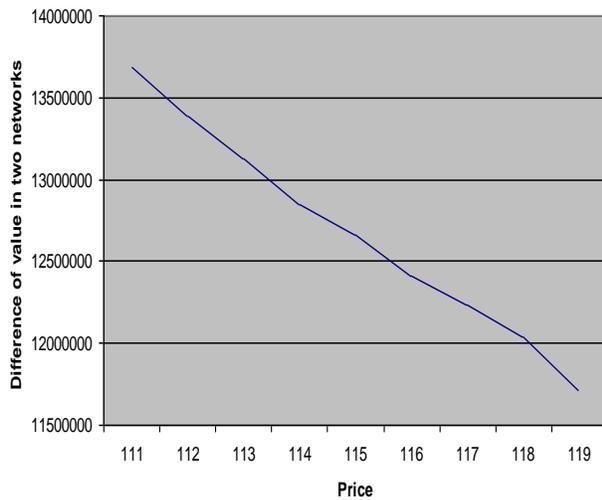


Figure 6. Dealers' difference of value in basic and transformed networks.

C. The Impact of New Entries

We call the network with the new group of dealers as the competitive network. We calculate values in the new scenario at mean repair price 216 which is the optimal price for the transformed network. We investigate the impact of the change of dealers letting the price unchanged so that the end customers are motivated to remain in the network. We show that dealers' value (31.527.812) is lower in the competitive network compared to the transformed one (35.481.031), since the new dealers' cost is higher due to the complementarities they offer. In addition, OEM's value increases (from 29.793.000.000 to 31.713.504.020) due to the increase of the service requests. The total value of the network increases from 32.190.040.300 to 32.792.529.000.

From the above we observe that a change in the network that improves its performance may affect positively some participants and negatively others. Naturally, dissatisfied participants abandon the network causing side effects to the others.

D. Participants' Equilibrium Strategies

We perform two experiments in order to investigate strategic interactions and determine equilibrium strategies of OEM and dealers. In the first experiment we calculate OEM's optimal profit rate at a given profit rate for the dealer. Simulations show that when the dealer increases its profit rate (e.g., from 6% to 10%), OEM's optimal choice is to decrease its optimal profit rate (from 24% to 21%). Conversely, if OEM increases its profit rates (e.g., from 14% to 21%), the dealer optimally decreases its profit rate (from 15% to 10%).

The second experiment calculates a set of equilibrium strategies for OEM and the dealer: at dealer's profit rate of

10% the optimal OEM's profit rate equals 21%. Conversely, at OEM's profit rate of 21% the optimal dealer's profit rate equals 10%.

VI. CONCLUSIONS

In this paper, we proposed a methodology that estimates value in service systems. We applied this methodology to a car repair service network. We run simulation experiments to maximize the value of each participant and the total value of the network. In addition, we defined suitable scenarios to study the internal relationships that are developed inside the service network. Finally, we examined the interactions between the participants inside the service network in order to determine their optimal choices.

Directions for future work include the study of competitive service networks that form oligopolies in order to increase value. Furthermore, additional work is needed on the estimation of value of intangible assets such as knowledge, sense of community, etc.

REFERENCES

- [1] J. Spohrer, P. Maglio, J. Bailey, and D. Gruhl, "Steps Towards a Science of Service Systems", *Computer* 40(1), pp. 71-77, 2007.
- [2] V. Allee, "The Future of Knowledge: Increasing Prosperity through Value Networks", Butterworth-Heinemann, Boston, 2002.
- [3] J. Gordijn and H. Akkermans, "Designing and Evaluating E-business Models," *IEEE Intelligent Systems* 16, No. 4, 11-17, 2001.
- [4] J. Sairamesh et al.: "Dealer collaboration: transforming the value chain through integration and relationships", *Proc. of International Conference on E-Commerce Technology (CEC'04)*, pp. 325-239, 2004.
- [5] N.S. Caswell, C. Nikolaou, J. Sairamesh, M. Bitsaki, G.D. Kouras, and G. Iacovidis, "Estimating value in service systems", *IBM System Journal*, vol. 47, nr. 1, pp. 87-100, 2008.
- [6] A. Biem and N. Caswell, "A value network model for strategic analysis", *Proceedings of the 41st Hawaii International Conference on System Sciences*, 2008.
- [7] V. Allee "Value Network Analysis and Value Conversion of Tangible and Intangible Assets" Published in *Journal of Intellectual Capital*, Volume 9, No. 1, 5-24, 2008.
- [8] C. Fornell, M.D. Johnson, E.W. Anderson, J. Cha, and Bryant B.E., "The American customer satisfaction index: Nature, purpose, and findings", *Journal of Marketing* 60, 7-18, 1996.
- [9] Jay W. Forrester, "Industrial Dynamics", Pegasus Communications, 1961.
- [10] <http://www.iseesystems.com/softwares/Business/IthinkSoftware.aspx>
- [11] <http://www.vensim.com> 02/12/2010
- [12] <http://www.powersim.com> 02/12/2010
- [13] J.D. Sterman, "Business Dynamics: Systems Thinking and Modeling for a Complex World", McGraw-Hill, 2000.
- [14] D. Fudenberg and J. Tirole, "Game Theory". MIT Press, 1991.

Model-Driven Dynamic Service Delivery in Mobility and Ambient Environment

Soumia Kessal/ Noémie Simoni/ Xiaofei Xiong/ Chunyang Yin

Department INFRES

TELECOM ParisTech - LTCI - UMR 5141 CNRS

Paris - France

{kessal, simoni}@telecom-paristech.fr/ {xiaofei.xiong, chunyang.yin}@gmail.com

Abstract— The rapid evolution of the new generation networks and services (NGN/NGS), particularly the converged infrastructures, raises the challenge of ensuring not only the Media Delivery but also the Service Delivery towards the user, regardless of his terminal, his access network, his preferences and his Quality of Service. The existing approaches for Media Delivery enable the improvement of the network in order to share and allocate the resources in the most optimal way. But the mobility and the heterogeneity of the user's ambient environment invoke more and more the complexity of the Service Delivery. In this paper, we propose a Dynamic Service Delivery process over Media Delivery in order to ensure the service continuity during the mobility or during all the other changes in the user's ambient environment. This Delivery process is driven by a model which represents the real world and differentiates the Service Delivery treated by the service platform from the Media Delivery treated by the network. The advantage of the proposed model is that it ensures the consistency within the deployment of user session and takes into account the End to End Quality of Service in the NGN/NGS context.

Keywords-NGN/NGS; Service Delivery; Model Driven; Dynamic Management; Mobility.

I. INTRODUCTION

With the new generation network (NGN), the user is facing a new environment composed by multiple access networks, core networks, service platforms, access terminals and even different operators. In such a varied environment, user hopes to change the terminal everywhere while maintaining the service continuity according to his preferences. A first response to this increasingly complex environment was the network convergence, which enables to provide the same services to the user through different networks. This convergence has thus allowed the passage from a vertical architecture (where the access network, the core network and the service platform are tightly coupled) to a horizontal architecture [1] [2] [3] (where the network and the service platform are on independent layers). If this architectural transformation succeeds, the objective of the trans-organization, the mobility and the personalization requirements, is fulfilled. However, in this service separated architecture, how to address the service management and keep the consistence with the delivery of the network transport service (media delivery)?

Service Overlay Network (SON) allows the access to the same types of services (Service Overlay) through different but provisioned networks (Network Convergence). Unfortunately, these architectures are still vertical from the point of view of resources reservation when providing the solutions for the improvement of the Media Delivery. Indeed, the infrastructures remain specific to certain types of services. As a result, there rests the passage from one type of service to another (for example, from a Telco service to a Web service), or the service composition will necessitate a change in the whole transport infrastructure.

Additionally, in this type of architecture, the maintenance of the service continuity during the mobility is only treated via the network (Media Delivery). But the End to End (E2E) Quality of Service (QoS) depends also on the QoS of Service Layer to achieve the demanded Service Delivery; moreover, the service layer can also provide complementary solutions. Indeed, depending on the current ambient network in which the user stays, we use an ubiquitous component of this area in order to have another network path and thus to meet the E2E QoS. Our main proposition called "Service Delivery" is based on this observation and takes account of mobility.

From this point of view, we replace the Service Overlay by a Service Network which enables service convergence and global dynamicity in the network layer and in the service layer as well. Architecturally, this service network is driven by a model and the Service Delivery enables this service network to facily provide user personalized services. This service network auto-manages itself independently of the transport infrastructures, and changes dynamically due to the mobility and the changes of environment in order to manage the real time service continuity during the delivery of the service.

The rest of the paper is organized as follows:

In Section 2, the related work shows the lacks of Media Delivery for the E2E service. After we introduce the models which are the base of our proposal (Section 3), we present in Section 4 our proposition in details. In the first place, we present as an example the Internet Protocol Television (IPTV) architecture to show the importance of the service level contribution. Afterwards, we present the important part of our contribution which allows the dynamic management to treat mobility. Then, in Section 5, we show how our proposal goes from the theory to the practice by two parts: for the Model-Driven of our delivery process we present how to anticipate

the degradations in order to maintain E2E QoS. On the other hand, we show how our model can help in the standardization; we propose the integration of our Service Delivery Process in the Tele Management Business Framework Process. Finally, we conclude the paper in Section 6.

II. RELATED WORK

To deliver the services to today's large numbers of mobile users although a user's ambient environment, the management of multiple composed services, the flexible provisioning of resources, and the guarantees of QoS remain as challenges. However, current solutions are not dynamic and are of the type Overlay as the following in SON [4]. The Service Overlay Network (SON) is proposed as an intermediate layer, which enables the access to the same services by converging the different networks. SON attempts to increase application adaptation and service reliabilities. In this section we analyze the existing optimization solution for SON in order to show the lacks.

A. Topology Re-configuration Policies in SON

As the reconfigurability is one of the most appealing features of SON, J. Fan et al. proposed optimal reconfiguration policies for the topology design of SON in order to improve the performance and to minimize the operation cost of dynamic routing [5].

As the communication requirement changes, the topology may need to be reconfigured with a lower-cost path. Comparing with other work, it is considered in a general case that the topology is not assumed as fully-meshed but as degree-bounded: the number of the neighbors of a node is limited. They analyzed and identified all types of cost during the phase of reconfiguration, observed the optimal reconfiguration of small systems, and then gave out the reconfiguration policies for large systems where the communication pattern is dynamic.

The policies are verified to have low cost when facing changes of communication requirements. As the nodes in the topology graph are assumed to be fixed and the numbers are limited, it is difficult to discover other service nodes in an unknown domain out of the existing SON.

The policies can resolve some mobility but in a limited range. It is not a trans-organizational solution.

B. An optimal routing based on service federation

M. Wang et al. proposed a distributed algorithm [6] of the service flow graph based on the service federation in order to optimize routing in SON. The service federation enables independent services to perform the tasks in either a sequential, parallel, or interleaved fashion by allowing parallel paths for a complex service.

To connect the service requirement and the overlay graph, they defined a service abstract graph, within which, a service abstract node is populated with the instances of the corresponding service. After applying the algorithm proposed by Z. Wang et al. [7] in the overlay graph, a service abstract

graph is constructed respecting to proposed strategies, and the same algorithm is applied again in the abstract graph, so that some links of the previous path maybe replaced by optimal ones. This service federation algorithm is proved to be able to construct service flow graphs with high-quality.

Although it can response dynamically in the service layer, but the algorithm is not flexible enough facing the ambient environment and service continuity during the process of the dynamic routing.

C. QoS-Assured Service Composition

X. Gu et al. presented an infrastructure for the efficient service composition, focusing on the assurance of QoS [8], which is in favor of the reliability within a SON. They propose a QoS-assured algorithm for the mapping from service template, which is mapped from service requirement, to an instantiated service path.

For the initiation of service composition, the basic algorithm generates the weighted candidate graph instead of searching the path in the first step, and then runs the Dijkstra algorithm to find the shortest path considering various QoS constraints. In order to consider each individual QoS constraint, an enhanced algorithm is proposed in additional, which can change dynamically the weight of different factors. At last, the dynamic service composition algorithms are proposed to deal with the recomposing of service path in a complete way or in a partial way, in the run time when outage or significant quality degradations occur.

The proposition is proved to be dynamic and to be able to provide both QoS assurances and load balancing for composed services in SON. As the algorithm considers the nodes and the links in the whole overlay topology, we still need to improve the flexibility and efficacy without increasing the complexity.

D. Auto-configuration in SON

R. Braynard et al. presents Opus, an overlay peer utility service which can automatically configure server network overlays by running instances of global resources and tracking their status, in order to improve the performance via the avoidance of unnecessary re-implementation [9].

A service overlay works as a "backbone", through which Opus can track and disseminate the information. For each individual application, Opus creates an overlay, and dynamically adapts the overlay topology according to the system load and network conditions. Opus determines the allocation of resources for competing applications, by maintaining a cellular structure, within which, a cell can be an entire Opus presence or a portion of it. To track the system characteristics, they applied to a generic communication library within Opus, some existing routing protocols which have the properties of aggregation, hierarchy and approximation. At last, to ensure reliability and availability, several approaches are pursued, such as restricted flooding and construction of disjoint paths.

It is a good idea to appoint a special management unit to implement the auto-configuration. However, it is too centralized. Can the management mechanism to be more

distributed which is self-manageable for each unit? And be more independent as possible to the infrastructures?

In conclusion, the Overlay Network does not have enough flexibility to meet the new NGN/NGS challenges.

III. BACKGROUND

In this part, we present the models on which our proposition is built and which gathers: the NLN (Node-Link-Network) meta-model [10] which represents all kind of component (§A), the QoS model to manage the behavior of the components (§B) and the Information Model which represent our knowledge base (§C).

A. NLN Meta-Model

In the delivery process, we must ask ourselves: what do we manage? Which type of resources do we want to provision? In order to be as generic as possible and to take into account all kinds of resources, we follow the NLN meta-model describing three types of objects: the Node which represents the component in each visibility level (Service, Network or Equipment) and which is responsible for specific treatment; the Link which represents the interaction between two Nodes, according to the service logic and the Network which represents all the Nodes and the Links offering a global service to each visibility level. This model enables the decomposition of the whole system in several abstractions' levels according to the service provided by each component, which allows a complete management of the system.

B. QoS model

In order to maintain the user Service Level Agreement (SLA), we must have an E2E QoS control during all the service provisioning and delivery processes. We propose a QoS agent in every component and the behavior of each component will be translated by QoS measurable parameters. QoS can be categorized according to four criteria [11]: Availability, Reliability, Delay, and Capacity.

These criteria are necessary and sufficient for self-control (IN/OUT Contract) because they are able to cover Fault, Configuration, Accounting, Performance, Security (FCAPS) framework proposed by the Telecommunications Management Network (TMN). Availability indicates the relation states for the configuration; Reliability depends on Fault; Delay and Capacity concern Performance and Accounting, and Security is considered as a service where we apply the QoS agent. This QoS model is instantiated on each visibility level and the aggregation of these parameters on all of the visibility levels ensures an E2E QoS. This QoS model allows managing the resources in use and their possible degradations.

C. Informational Model (IM)

The IM we propose is generic and abstract to describe any resource. This IM, as our knowledge base, is independent from application. It is directly related to the events that occur in the

real network. The inference rules (implications of each event) have been defined in [12] [13].

In the next section, we present our proposal based on the model which allows the same architecture for the Service Delivery and Media Delivery.

IV. PROPOSITION

Our proposal takes into account the usage in the NGN/NGS context, essentially in a mobile and user-centric oriented context. Indeed, between the user's demanded service and its dynamic utilization in this context, it is necessary not only to provision the service at the moment of its activation, but to re-provision continuously and dynamically according to the user's location and the terminal in use. In addition, the service presented on a catalogue (for example, the Triple Play) incites, in terms of resources, several service elements (Voice, Data, TV), which need to be integrated in our dynamic management (Provisioning, Re-provisioning and Assurance). In this section, we firstly show the necessity of Service Delivery complementary to Media Delivery, by presenting in the IPTV architecture the application services and the network services, then explain the importance of Service Delivery in the service layer (§A). In (§B), we analyze the needs impacted by all types of mobility (User, Terminal, Network and Service), and the ambient networks offering ubiquitous services in order to meet the NGN/NGS context. Then, in order to manage the dynamicity during the usage of service (during the movement), we propose to follow the traceability of the dynamic session basing on the NLN model through the different layers (Service, Network and Equipment) (§C). In (§D), we present the process of E2E Service Delivery, which takes charge in the usage and complements the Media Delivery. At last, we detail the models (Virtual Service Community and Ubiquitous Service Element) which manage the mobility in the delivery process (§E).

A. What is the Service Delivery

In this section, we explain the Service Delivery and its contribution relative to the Media Delivery.

During the phase of provisioning, we need two to provide a service requested by a user. The first is the selection of the requested service according to the user's preferences, his location and his terminal. The second is the selection of the network that can satisfy the requested QoS. Within the ambient context, the user moves while having the continuity of service. Current Media Delivery solutions treat these requirements by shifting the access point and the corresponding supporting services in the network layer. The QoS is obliged to be recalculated and sometimes the delivery is interrupted, although we are now facing an ubiquitous context, i.e. we are already able to predict towards which pervasive service the user can be served while maintaining the desired QoS. That's why we believe that a higher level concept of Service Delivery is needed to address the dynamicity during service usage. We benefit from the ubiquitous context in the level of the service platforms. The service continuity can be obtained by dynamically managing

the Service Delivery. During this dynamic management, we can adapt to the change of the user's location and to the QoS degradation, and can always conform user's preferences. Once the Service Delivery is in place, the network layer follows the solution and provisions the resources to provide QoS continuity.

In the following, we present the IPTV architecture and show the need of the Service Delivery, which complements the Media Delivery.

The IPTV service [14] is becoming more and more popular among telecommunications companies because it promises to deliver TV programs at anytime and anywhere. Based on IP protocol, IPTV features advantages like bandwidth efficiency and ease of management.

We can identify three main parts of the overall architecture (Figure 1):

Transport Functions: RACS, NASS and Transport Processing.

IPTV service control and applications functions: Service Selection Function (SSF), Service Delivery Function (SDF) and Service Control Function (SCF).

Media Delivery, Distribution and Storage functions: Media Delivery Function (MDF) and Media Control Function (MCF). The MCF has three interfaces: one with application service (SCF), one with network service (MDF) and another with the User Equipment (UE) to control the remote.

In this architecture, there are application services (offered by SSF, SDF, SCF or MCF) and network services (offered by MCF or MDF). The application services identify the flow-generator services which must be transmitted through the network. Such services might change during the transmission.

NASS and RACS, which manage the Policy Enforcement Point (PEP) as policy control to ensure the QoS.

These two types of services work in an independent way, which separate the service from the network. However, the actual solution for the terminal mobility is oriented to the access network, i.e., the access part detects the new location and the Media Delivery part will not rearrange another delivery flow until the transport deviation is finished. As a result, the user service is not continuous due to the non seamless of the service handover. If we can realize the Service Delivery which aims at delivering complexes services (as IPTV) to today's large numbers of nomadic users through out a user's ambient environment, we can thus achieve the service continuity with the user desired QoS.

B. Requirements analysis of Service Delivery

In order to realize the Service Delivery concept, we need to focus on the mobility management, the flexible provisioning of resources and the guarantees of service continuity and dynamic E2E QoS during the mobility.

First of all, the ambient environment is actually a very heterogeneous environment due to different access technologies, services and networks environments. The competition and the cooperation of various market players are facilitated by defining interfaces, which allow the instant negotiation of agreements. This new environment enables services to be pervasive, i.e. the same type of services from different suppliers are all visible to a user.

As a result, this new context challenges us:

- To offer the ubiquitous services. This aims to provide the same services in the user's ambient environment during the movement.
- To take into account the user's preferences. This might influence the choice of service, of operator, of access network, of equipment or of price rate against another.
- To provide personalization by composing heterogeneous services during a session. This requires horizontal management to be independent to the different architectural layers (Service, Network and Equipment).

Secondly, another major challenge rose is the mobility management during the user mobile session. We summarize four types of mobility that might occur during the session facing the ambient environment.

The terminal mobility refers to a terminal which moves through different access points while maintaining its connectivity.

The network mobility concerns the movement of the infrastructure of the transport support.

The service mobility refers to the ability of service to be transferred from one machine to another so that the user can use the service independently to the terminal or network in use.

The user mobility concerns that the terminals are switched by a user, which requires the adaptation of service on the new terminal.

This mobile context imposes us:

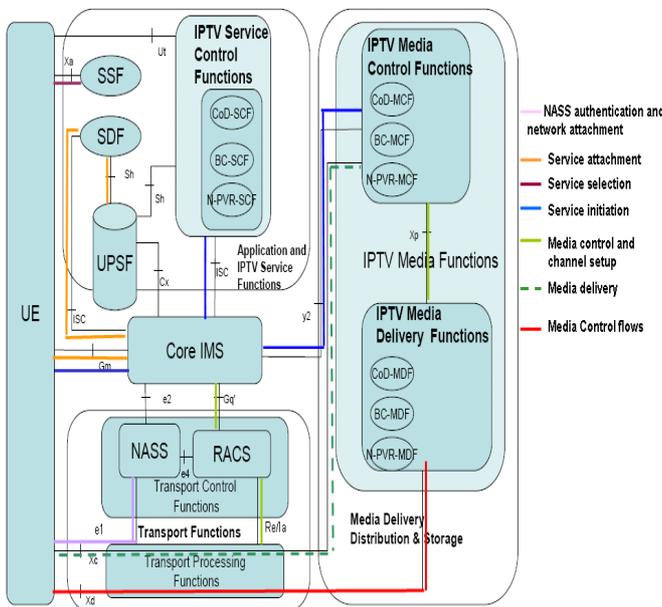


Figure 1 Applicative Services and Network Services in the IPTV Architecture

The MCF and MDF are concerned with the media transport via all the network components with the help of

- To guarantee the continuity of service all through these types of mobility. This results in the session mobility. The session mobility concerns the dynamic management of the user session due to its E2E uniqueness in the real time. In fact, each type of mobility belongs to an architectural layer (Equipment, Network or Service), and is not able to ensure the E2E QoS by itself. It is the session management in the real time that maintains this E2E QoS.

- To ensure the Re-Provisioning of resources during the movements. The new usage induces the consideration of the preferences and the mobility of the user.

Our goal is to meet these needs by proposing a process of E2E Service Delivery driven by mentioned NLN and QoS models, which we detail in the next section. The Service Delivery takes in charge of the changes impacted by all types of mobility on the service platform in the ambient environment.

C. Model-driven Service Delivery

Our model thus consists of several concepts to response the requirements identified in §B:

- The service element (SE): we model it ubiquitous and mutualisable. Its ubiquity answers especially to the ambient environment. The service elements are equivalent in function and in QoS and installed on the same platform or on different platforms. The characteristic of mutualisation enables the sharability between multiple users, which allows, as consequence, different users' preferences and a more dynamic and efficient participation of the service composition.

- The model NLN presents in the background: it allows us to model the heterogeneous real world via the nodes and the links. The network of nodes and links of the same nature constitutes a virtual network. Taking advantage of these virtual networks, the required horizontal management obtains its architectural base.

- The concept VPxN: allows the provisioning of resources in each layer of visibility. We propose to model each architectural layer by a Virtual Private x Network: VPxN (x=Equipment, Network or Service), which auto-manages itself dynamically during the movements. The VPxN allows re-provisioning the resources to the user according to the changes (mobility, preferences, degradation) for the purpose of the E2E QoS continuity. Thanks to the VPxN, during the mobility we can offer the user an ambient environment (all the resources of equipments, networks and services which enable the session continuity in the new location).

- The concept Virtual x Community (VxC): allows creating communities of equivalent resources in function and in QoS at each visibility layer. Thanks to VxC, during mobility we can anticipate the degradations by replacing the degraded resource by another equivalent one.

As a result of the application of these models, we have our global architecture as the following:

The VPEN, which regroups all the ambient equipments of the user: terminals, networks equipment and servers.

The VPCN, which consists of access and core networks. It regroups all the ambient connectivity of the user: Wifi access, BTS access, ADSL access, IMS core, etc.

The VPSN, which makes up a logic network of all the accessible services to the user. These services are managed in a horizontal way (independent from any particular network infrastructure).

Each visibility layer (VPEN, VPCN, and VPSN) is dynamic and self-manages in a horizontal way. Thus it is the session management which takes into account the different layers of visibility for the calculation of the E2E QoS. For example, given the user's preference of terminal, it is decided which ubiquitous service element to use, through which network, in order to guarantee the QoS continuity. Finally, to meet the needs of the dynamic management of the service elements during the mobility, the VxC monitors the comportment of each resource and proceeds to its replacement in each VPxN. The replacement is effected following any types of changes in order to maintain the continuity of the QoS.

Therefore, we have three layers of visibility in the global architecture: equipment, network (access and core) and service platform (Figure 2).

In the equipment layer, we have represented all the equipments in the PAN of the user, the equipments of his access network and of his core network, and the servers on which all the service elements accessible to the user are installed.

In the network layer we have represented the access network and the core network chosen for his session.

In the service layer we have represented the logic network (VPSN) of all the services to which the user has the right to access.

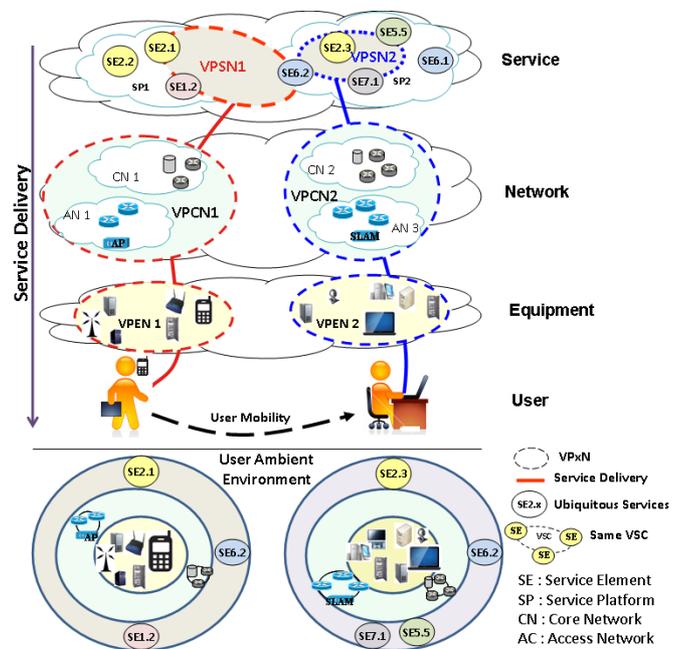


Figure 2 Service Delivery with QoS Continuity according to Mobility, User's preferences and Ambient Environment

From the VPxNs, in a given location, we can thus give the whole ambient environment of the user. For example, the user in his first location has chosen his cell phone as terminal during his session, he passes through the selected access network and the core network (VPCN1), and he may use the service elements (SE2.1, SE1.2 and SE6.2) during his session. All the service elements are ubiquitous (of the same color in the Figure 2) and distributed on two different platforms (SP1 and SP2).

During this session, user mobility occurs, i.e., he changes the terminal from a cell phone to a computer, which will cause a change in the access and core networks (VPCN2 in the Figure 2). This may require a change of a service element to adapted to the new terminal, or maintaining the same service element on another platform while changing the network or the terminal. In the example, an ubiquitous service (SE2.3) on another platform (SP2) replaces SE2.1. This change may be treated with during the transaction of the calculation of the E2E QoS, which selects this ubiquitous component in order to guarantee the demanded QoS. After the mobility, the user has still two new elements (SE7.1 and SE5.5) adapted to his terminal, which replace the previous element (SE1.2). The VSC management takes in charge of the efficient replacement of the service elements by anticipation.

We present the E2E Service Delivery Process in the following part.

D. Service Delivery Process

When the user orders the services, the service provider firstly provisions the resources to meet his demand. In our proposed process, this step is the creation of a “Pre-VPSN”. The Pre-VPSN will contain all the services subscribed by the user. The service provider will add the necessary services for this user. At the session initiation, following his service logic, the user constitutes the services that might be used all through the session which will be the “VPSN”. Each service element in the VPSN belongs to a VSC which manages it dynamically during the mobility. During usage, the service elements requested will be reserved and constitute a Transaction “Active VPSN”. We have illustrated the different steps (Pre-VPSN, VPSN and Transaction) in the Figure 3.

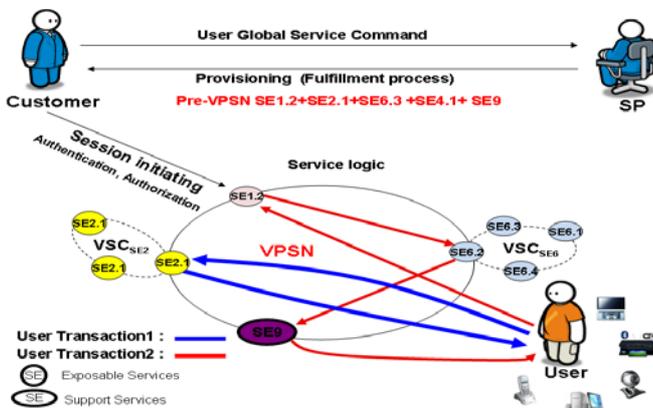


Figure 3 Pre-VPSN, VPSN and Transaction processes

In the following, we detail the processes which manage the different steps.

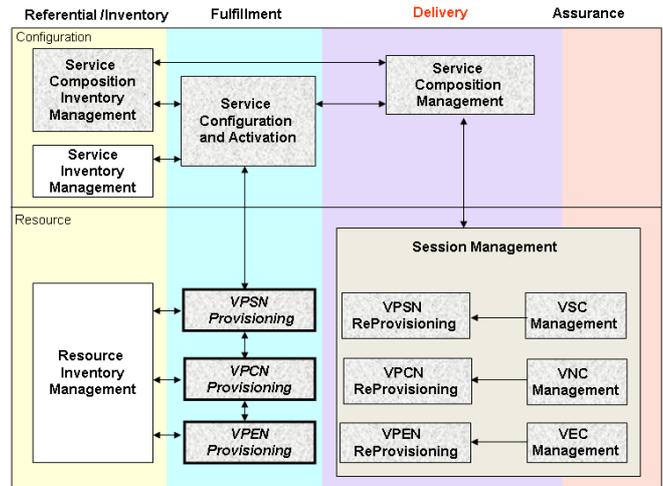


Figure 4 Service Delivery Process and his interaction with Fulfillment and Assurance

For the Order phase, we have the Fulfillment process (Figure 4). We have two parts: Configuration and Resource.

- The Configuration part takes in charge of service elements configuration and activation according to client order (Service Configuration and Activation process). This process communicates with the process (Service Inventory Management) in order to have the available services and their configurations. The Configuration part sends a request to the Resource part (OSS) to provision the demand.
- In the Resource part, we have the provisioning of three visibility layers [15], thus the process (VPSN Provisioning) provision the service resource. This process sends a request to the process (VPCN Provisioning) in order to reserve the network resources selected for the required services. Then the request is sent to the process (VPEN Provisioning) to reserve the user equipments, network equipments and service equipments selected for the demand. These three processes communicate with the data base “Resource Inventory Management” to find the available resources, their QoS, their addresses, and etc.

For the Usage phase, we have the Service Delivery process (Figure 4). This process manages mobility and changes in the user ambient environment. There are two parts in the process:

- In the Configuration part, we have “Service Composition Management” which takes in charge of the management of the services subscribed by the user. These services represent a user's VPSN at his session initiation. The VPSN contains all the service elements that the user has chosen during his session. The information of these elements is stored in the proposed data base: Service Composition Inventory (Figure 5).

User	Pre-VPSN at order	VPSN at the session opening	Active VPSN at usage (transaction)
User 1	SE1	Exposable Services	SE1
	SE4		SE4
	SE2	non Exposable Services	SE8
	SE8		SE8
User2	SE1	Exposable Services	SE3
	SE3		SE3
	SE5	non Exposable Services	SE5
User3	SE5	Exposable Service	SE5
	SE7		SE7
	SE8'	non Exposable Services	SE8'
	SE8'		SE8'

Figure 5 Pre-VPSN and VPSN example in Inventory

- In the Resource part, we have the processes of Re-provisioning of VPSN, VPCN and VPEN. Each VPxN auto-manages itself in a dynamic way and re-provisions the resources during mobility, preference changes or resource degradation. It is the session which deals with the aggregation of the three layers in order to handle the demanded service with QoS continuity. The "Session Management" thus regroups the three layers of re-provisioning and community management which auto-manage themselves ensuring session and QoS continuity.

For the Assurance phase, we have the Assurance process (Figure 4), where we have the management of VEC, VNC and VSC. These components monitor the behavior of the resources (the behavior results in QoS). Each resource announces its conformity or its non conformity to the contract with the members of its community. In the case of a non conformity thus of a QoS degradation, the resource is replaced in each VPSN, VPCN or VPEN associated by a resource functionally and QoS equivalent of the appropriate community.

In the next section, we detail the mobility management treated by VSC and ubiquitous service elements.

E. Taken into account mobility (VSC and Ubiquitous SE)

The VSC we proposed are built at the deployment phase. When a service is deployed in a location it will search the ubiquitous service elements that are identical to him in terms of QoS and functionality, to represent the service community. This community aim to self-manage faults or QoS degradation. The goal to have these communities is that, when a service element is degraded and filled not its contract it will find in its location the nearest and similar service element which can replace it in the VPSN. Thus with this concept we can anticipate degradations in order to allow service continuity.

During mobility, let us suppose that we have degradation in a service element (e.g., SE1.2); degradation is alerted by SE1.2 community. This community will carry out the replacement of this service element by another equivalent in the VPSN and will thus allow service continuity. But before replacing this element, we first calculate the QoS of the link (transport layer) between this new service element to replace (e.g. SE1.3) and the next and previous service element in the VPSN (e.g. SE6.5 and SE7.3).

From the service element QoS table (TABLE I) on Server1, Server2 and Server3, we have the QoS of SE1.2, SE1.3, SE6.5 and SE7.3.

TABLE I. QoS TABLE OF SERVICE LAYER: SERVER1, SERVER2 AND SERVER3

SE	Server 1	SE	Server 2	SE	Server 3
SE1.2	QoS _{SE1.2}	SE1.3	QoS _{SE1.3}	SE1.5	QoS _{SE1.5}
SE2.3	QoS _{SE2.3}	SE6.3	QoS _{SE6.2}	SE7.3	QoS _{SE7.3}
SE7.5	QoS _{SE7.5}	SE6.5	QoS _{SE6.9}	SE6.5	QoS _{SE6.5}

From the Transport Layer QoS table (TABLE II), we have the QoS of all the links (networks) possible between Server1 and Server2, and between Server3 and Server2. Thus to replace SE1.2 installed on Server1 by SE1.3 installed on Server2, we have {QoS_{SE1.2}, QoS_{(1-3)'}, QoS_{(1-3)''}} and after replacement we have {QoS_{SE1.3}, QoS₍₂₋₃₎, QoS_{(2-3)''}}, which allows maintaining the E2E QoS during the current transaction.

TABLE II. QoS TABLE OF TRANSPORT LAYER: SERVER1 AND SERVER2

	Server 2	Server 3	Server 4		Server 1	Server 3	Server 4
Link	QoS ₍₁₋₂₎	QoS ₍₁₋₃₎	QoS ₍₁₋₄₎	Link	QoS ₍₂₋₁₎	QoS ₍₂₋₃₎	QoS ₍₂₋₄₎
Link'	QoS _{(1-2)'}	QoS _{(1-3)'}	QoS _{(1-4)'}	Link'	QoS _{(2-1)'}	QoS _{(2-3)'}	QoS _{(2-4)'}
Link''	QoS _{(1-2)''}	QoS _{(1-3)''}	QoS _{(1-4)''}	Link''	QoS _{(2-1)''}	QoS _{(2-3)''}	QoS _{(2-4)''}
Link'''	QoS _{(1-2)'''}	QoS _{(1-3)'''}	QoS _{(1-4)'''}	Link'''	QoS _{(2-1)'''}	QoS _{(2-3)'''}	QoS _{(2-4)'''}

V. FROM THEORY TO PRACTICE

In this section, we show the feasibility of our proposal through two parts: For the Model-Driven of our delivery process we will present in (§A) the implementation of the VSC by JXTA [16] which allows us to anticipate and deal with the changes of user's ambient environment in order to maintain E2E QoS. To verify the consequence of considering the mobility during usage via the service management processes, we show in (§B) how our model supports the standards of Tele Management Forum.

A. VSC Implementation in JXTA

As we have explained, the VxC takes in charge of the maintenance of the resources (Equipment Element, Network Element, Service Element) in each layer, for example, a VSC in the service layer contains all Service Elements equivalent both in function and in QoS in the service network. In the first place, each SE needs to publish its function and QoS values to the others and receive the information from the others, so that VSCs can be constructed. Secondly, a VSC should have the ability to be aware of the change of its SE and to discover the unknown SE which can be added into the current VSC.

In our feasibility, we use the JXTA platform, which includes a set of open, generalized peer-to-peer (P2P) protocols. With the JXTA technology, all the elements can be regarded as peers, and can communicate and collaborate with each other in a P2P manner by publishing an advertisement with the form of language-neutral meta-data structures represented as XML documents. Within a sub network, the

peers can communicate directly with each other by Peer Advertisement. In JXTA, a Rendezvous peer can maintain global advertisement indexes of the peers that register to it in a sub network, and exchange information with other Rendezvous peers of the other sub networks. Therefore the peers can get the information of the peers in another sub network by communicating with its Rendezvous peer via Pipe Advertisement.

In our case, each SE publishes an advertisement including its characteristics, such as supported function, designed QoS, current location and etc. At the same time, each SE listens the advertisements from other peers in its neighborhood, and keep in a local table those whose function is the same with itself, basing on which a VSC is constructed.

```
<?xml version="1.0"?>
<!DOCTYPE jxta:PeerAdvertisement>
<jxta:PeerAdvertisement xmlns:jxta="http://jxta.org">
<PeerId>urn:jxta:uuid-280984B767FF4B80980DE4586287229305</PeerId>
<Name>SE1</Name>
<Layer>Service</Layer>
<Function>IPTVMediaControlFunction</Function>
<QoSCapabilities>
  <Delay>15ms</Delay>
  <Reliability>98%</Reliability>
  <Availability>99%</Availability>
  <Capacity>2.5Mbps</Capacity>
</QoSCapabilities>
<Location>PointA</Location>
</jxta:PeerAdvertisement>
```

Figure 6 Peer Advertisement published by a SE

In Figure 7, before the VSC is constructed, the two elements with the same shape and the same color are two elements of the function MCF with the same designed QoS. They need to find each other and add each other to its own VSC. Each publishes a peer advertisement within the sub network shown in Figure 6, and a pipe advertisement (same contents as in the peer advertisement) to its Rendezvous so that the peers in another sub network can also receive its information. When the convergence of this phase is finished, the SE of IPTV MCF in platform 1 will know that in platform 2 there is another SE of MCF which has the same QoS, and adds it to its table. Therefore, these two SEs make up a VSC of MCF (Figure 7). After being successfully created, the VSC will maintain its SEs by the current QoS values in the real time in order to help VpxN to adapt to the mobility and the changes of the user's ambient environment

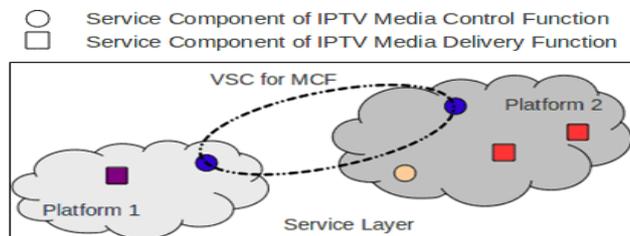


Figure 7 Creation of VSC for MCF

B. Valorisation on Standards

In this section, we show how our model helps in the standards. The Tele Management Forum (TMF), one of the most general forums in the domain of telecommunication management, has proposed a framework to define a complete management environment. The TMF includes the Business Process Framework (eTOM) [17] which can help service providers to define and describe the business and functional processes they use, to understand the links between these processes, and to identify the working interfaces and external entities, in order to structure the basic management Processes.

This management processes today are disjointed: the service is first provisioned following the customer's order, then used and assured, then billed [18]. However the NGS requires interactions between service logic during the usage. This interaction must be dynamic and takes into account the NGN/NGS challenges (e.g., User preferences, mobility and ubiquity). To meet these new paradigms we proposed to incorporate a usage process (Service Delivery) in the eTOM cartography. This process responds to NGN/NGS and interacts dynamically with the eTOM operations processes. In order to have coherence in this interaction we are based on an abstraction model of the real world for these processes (Fulfillment, Assurance, Billing and Delivery). In this model we have models each resource visibility level as a virtual private network (VPSN, VPCN and VPEN). We have also the VxC model that monitors the behavior of each resource and replace it by another ubiquitous resource in the VPxN in the case of QoS degradation. This can have a generic reference model for the NGN/NGS service management.

But before the operations processes, we have the Strategy, Development and Deployment processes. In order to have business continuity from strategy to operations, we have proposed the same model that allows consistency between the different processes. Thus these processes exchanges information dynamically. In order to be trans-organizational and to have the possibility to deploy services that are not developed by the provider himself but bought from other suppliers, we proposed a separation between the deployment and the development processes in the eTOM cartography. We also proposed new processes to take into account when dealing with the NGN/NGS context. An example of these processes is the "Community Creation" (see Figure 8, Deployment Process) who creates and manages communities in the different visibility level (VSC, VNC and VEC). These communities are created at the Deployment phase and managed in the operation phase. Thus, our model facilitates the relationship between the SIP processes and Operation processes.

VI. CONCLUSION AND FUTURE WORK

In this paper, we proposed a dynamic Model-driven Service Delivery in an NGN/NGS context with mobility and ambient environment. The proposed Service Delivery focuses on the service layer, where we use the solutions provided by the network (Media Delivery) to guarantee the service continuity.

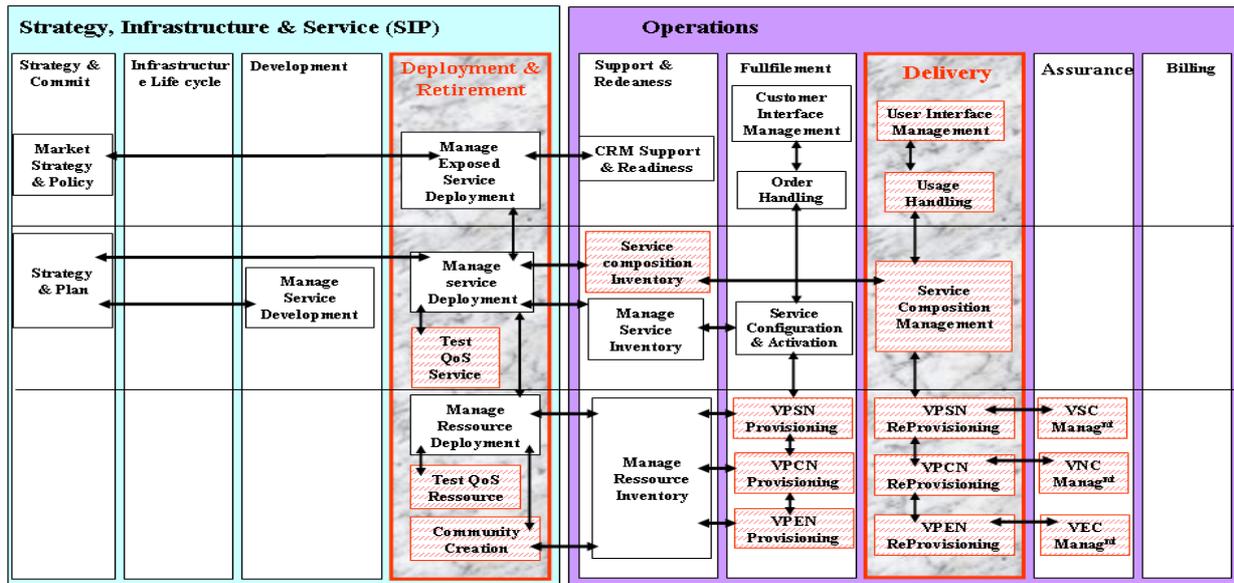


Figure 8 'Service Delivery' process integrated in the eTOM cartography.

And also the solutions provided by the service platform (ubiquity, mutuality, composition and independence of services in the transport layer) to ensure QoS and service continuity during the session. The proposed model (SE, NLN, VPxN and VxC) enables the coherence among the different layers. The VSC enables the anticipation of degradation in order to ensure the QoS continuity. The implementation of VSCs in JXTA, allows us to verify the feasibility of our QoS Management by VSCs in P2P. The VSC and Ambient Environment concept that we have defined to resolve management problems allows taking account into the mobility between different ambient areas. It allows us to make a "Semantic Handover" on other triggers criteria that the QoS management. Thus as a future work, we will finalize the Semantic Handover automate according to different criteria.

REFERENCES

[1] ETSI DTR/USER-00029-1 (TR 102 805-1), "User Group; End-to-end QoS management at the Network Interfaces; Part 1: User's E2E QoS - Analysis of the NGN interfaces (user case) , Control plan solution: QoS signalling", 2009.

[2] ETSI DTR/USER-00029-2 (TR 102 805-2), "User Group; End-to-end QoS management at the Network Interfaces; Part 2: Control and management planes solution - QoS continuity, Management plan solution: QoS Inter-working", 2009.

[3] ETSI DTR/USER-00029-3 (TR 102 805-3), "User Group, End-to-end QoS management at the Network Interfaces; Part 3: QoS informational structure", 2010.

[4] J.W. Kim, S.W. Han, D.H. Yi, N. Kim, and C.C.J. Kuo, "Media-Oriented Service Composition with Service Overlay Networks: Challenges, Approaches and Future Trends", Journal of Communications, Vol. 5, No .5. (2010), pp. 374-389, May 2010. doi:10.4304/jcm.5.5.374-389.

[5] J. Fan and M.H. Ammar, "Dynamic Topology Configuration in Service Overlay Networks: A Study of Reconfiguration Policies", - Proc. IEEE INFOCOM, pp. 1-12, Spain, 2007. doi: 10.1109/INFOCOM.2006.139.

[6] M Wang, B Li, and Z Li, "sFlow: Towards Resource-Efficient and Agile Service Federation in Service Overlay Networks", 24th IEEE

International Conference on Distributed Computing Systems (ICDCS'04), pp. 628 – 635, 24-26 Feb. 2005, Japan. doi:10.1109/ICDCS.2004.1281630.

[7] Z. Wang and J. Crowcroft, "Quality-of-Service Routing for Supporting Multimedia Applications", IEEE Journal on In Selected Areas in Communications, IEEE Journal on, Vol. 14, No. 7. (1996), pp. 1228-1234. August 2002. doi:10.1109/49.536364 Key: citeulike:2831634.

[8] X. Gu, K. Nahrstedt, R.N. Chang, and C. Ward, "QoS-Assured Service Composition in Managed Service Overlay Networks", Proceedings of the 23rd International Conference on Distributed Computing Systems (ICDCS '03), June 2003, pp. 194 – 201, doi: 10.1109/ICDCS.2003.1203466.

[9] R. Braynard, D. Kostic, A. Rodriguez, J. Chase, and A. Vahdat, "Opus: an Overlay Peer Utility Service", Proceedings of the 5th International Conference on Open Architectures and Network Programming (OPENARCH 2002), pp. 167 – 178, August 2002. doi: 10.1109/OPNARC.2002.1019237.

[10] N. Simoni and S. Znaty, Book . Gestion de réseau et de service: similitude des concepts, spécificité des solutions. ISBN: 2225829802.

[11] N. Simoni, B. Mathieu, C. Yin, and M. Song, " Autogestion de service par la QoS dans un Réseau Overlay ", GRES.07, Nov. 2007, Tunisia.

[12] N. Ornelas, N. Simoni, K. Chen, and A. Boutignon, "VPIN: User-session knowledge base for self-management of ambient networks", UBICOMM'08, pp. 122-127, Oct. 2008, Spain. doi: 10.1109/UBICOMM.2008.71.

[13] N.Ornelas, N.Simoni, C.Yin, and A.Boutignon, "VPIN: An event based knowledge inference for a user centric information system", Journal On Advances in Internet Technology, Vol. 2, N° 1, pp. 29-44, Sept. 2009.

[14] ETSI TS 182 027, V2.0.0, "Telecommunications and internet converged services and protocols for advanced networking (TISPAN), IPTV Architecture", Technical specification, 2008.

[15] S. Kessal and N.Simoni, "Service provisioning oriented QoS in NGN/NGS context", unpublished.

[16] Book.JXTA Java™Standard edition v2.5: Programmers guide, September 2007.

[17] TMF-GB921, "Business Process Framework (eTOM) : Business process framework", release 8.0.

[18] TMF-GB921, "Business Process Framework (eTOM) : Representative process flows", release 8.0.

Mechanism Design for Designing Annotation Tools

Roberta Cuel, Oksana Tokarchuk, Marco Zamarian

Dept. of Computer and Management Sciences

University of Trento

Trento, Italy

Email: (roberta.cuel, oksana.tokarchuk, marco.zamarian)@unitn.it

Abstract—The Web 2.0 is increasingly considered as a phenomenon that affects the way people interact, search, post and share information on the Internet. Namely it affects the daily life of any Web user, expert or company that works on the network daily. One of the dominant traits of Web 2.0 applications is the capability of co-opting end-users in endeavors which traditionally have been considered as top-down activities and exploiting user-based networks. Through Web 2.0 applications users add content and annotation in order to describe and share pictures, videos, files, etc. Apart from some of the most well known applications (e.g., Facebook, Twitter, Flickr and the like), many Web 2.0 tools are not good at attracting a critical mass of individuals. In fact, many studies have shown that a common outcome for online communities is an 'onion' structure where only a few core individuals actively participate discussing and contributing to the common content, whereas others are considered as peripheral users who observe the community and simply use the content. Participation and willingness to contribute, thus, become two of the critical issues that companies and software developers should take into account when creating Web 2.0 applications. In other words, we claim that understanding and analyzing appropriate sets of incentives that might motivate users to contribute are critical steps in the design of Web 2.0 applications. In this paper, we describe how theories and techniques that are well known and used by scientists in economics and management studies can be used to develop incentive-compatible Web 2.0 tools. Specifically, we will provide an example of an application of mechanism design and applied experimental economics in the development of an annotation tool.

Keywords—*Incentives; Mechanism Design; Web 2.0; Content Creation and Annotation.*

I. INTRODUCTION

The collaborative, social way of generating, organizing, and managing knowledge has been growingly considered as a trigger of creativity and innovation in several applied fields. This phenomenon has been heavily shaped by the advent of social based technologies such as grid computing, peer to peer file sharing, collaborative authorship of digital content, social networks, and, more in general, Web 2.0 applications [7], [18], [28].

In this scenario, the increasing popularity of Web 2.0 applications dramatically changes the way people interact, have fun, communicate and consume. The common trait of Web 2.0 applications is the empowerment of end-users by co-opting individuals in endeavors which traditionally have

been considered as top-down activities, and the exploitation of user-based networks of relationships [20], [23]. A glaring example of such a trend is the new emerging crowd-sourcing phenomenon [12], [14], [17]. Nowadays there are many Web 2.0 applications that can be considered as concrete examples of the crowdsourcing phenomenon: SourceForge (sourceforge.net), Wikipedia (en.wikipedia.org), Galaxy Zoo (www.galaxyzoo.org), Crowdfunder (crowdfunder.com), Innocentive (www2.innocentive.com). All these applications take advantage of communities of experts and/or users who proactively contribute to the creation of a common good. In economics, a good or service is called a 'common good' [21] when it has contemporaneously the properties of non-exclusivity, and non-rivalry. Non-exclusivity means that it is impossible or costly to exclude any person from the use of the public good. Non-rivalry means that each person can consume simultaneously the public good, without constraining the use for others. In all of the above-mentioned examples, users act according to their own personal purposes and, at the same time, provide information and knowledge that others can share and, in turn, use.

Many studies on online communities and peer to peer collaboration identified the motivations that drive people to participate in a rather large set of heterogeneous and context dependent elements [2], [13], [15], [25], [26], [27], [29] (e.g., reputation, altruism, competition, self-esteem, money, reciprocity, fun, etc.).

Despite these many motives to contribute and the popularity of many Web 2.0 applications, some studies observe how a significant number of online communities and social network applications fail because of under-contribution of participants. For instance, an analysis of the P2P file sharing site Gnutella show that in 2000, only 25% of users shared 98% of the content while 66% of users shared nothing [1].

More specifically, insights concentrating on the patterns that characterize annotation efforts in Web 2.0 communities found that annotations are characterized by power law distributions, both in the relationship between number of tags and number of posts [5] and number of tags and number of contributors [10], indicating that few people contribute disproportionately more than others.

It follows that not all Web 2.0 tools - and specifically the ones that take advantage of social networks - can become

killer applications *à la* Facebook or Twitter. This is due to the fact that users are not motivated enough to spend time interacting and contributing to a common good. Thus, motivating users to contribute to this kind of collective effort is essential to reach critical mass and ensure a sustainable growth for these crowdsourcing applications.

Since we consider the whole software development process very relevant, we claim that success of an online community-based application requires a blend of well-designed software (i.e., usability) and carefully crafted policies aimed at achieving participation. In this paper, we focus our attention on the social aspects of software development and deployment and we offer an example on how these aspects can be incorporated into the development of annotation and semantic content creation tools. Specifically we will adopt a set of methods and techniques, often referred to as 'mechanism design' in the field of economics, that can be used to develop incentives which can be embedded into online applications. This choice does not reflect a disregard for the technical aspects of software design, but is meant to underline features of the process that are oftentimes neglected within the community of developers.

We will focus on the so called sociability design and in particular on the first two phases of the software development process. These are the analysis of the use scenario prior to application design and the fine tuning process of the incentive structure. These phases should be seen as a continuous improvement process that enables designers to adjust the software according to the social needs emerging from the users' experiences.

The paper is structured as follows: Section 2 describes some basic notions of motivations and mechanism design, Section 3 sketches out some techniques for analysis and design of incentivized applications, Section 4 describes the analysis of an annotation prototype, and finally Section 5 draws our conclusions.

II. BASIC NOTIONS AND DEFINITIONS

Bouman et al. [4] argue that designers of social software have to design software and carefully craft social policies such as: enabling practice, mimicking reality, building identity and actualizing self. In order to effectively design social software we focus our analysis on motivations in the context of Web 2.0 and on tools that enable us to identify key incentives that can be embedded in the software.

A. Motivations

Several studies on the motivation to participate in knowledge sharing indicate that people participate because they want to be part of a 'community', and engage in the exchange of ideas and solutions [26]. Similarly, Forte and Bruckman find that peer recognition plays a role in Wikipedia which is similar to the dynamics shaping up scientific collaboration [9]. Wang and Fesenmaier [24]

demonstrate that efficacy is a major factor affecting members' active contribution to online communities. The study also indicates that the possibility of future reciprocation (expectancy) is another major motivation driving an individual's contribution. Beenen et al. show that challenging goals are powerful motivators of online contributions [2], while Wasko and Faraj found that people contribute when they perceive that this enhances their professional reputation [27]. Kuznetsov argues that the motivations of Wikipedians to contribute are grounded in the values of reputation, community, reciprocity, altruism and autonomy [15], [25]. Wiertz and de Ruitter found that a customer's online interaction propensity, commitment to the community, and the informational value s/he perceives in the community are the strongest drivers of knowledge contribution [29]. Bock and colleagues suggest the provision of appropriate feedback to employees engaged in (or not engaged in) knowledge sharing [3]. These actions follow from two considerations. On the one hand, they leverage on the importance of pressure exerted from a person's reference group (e.g., peers, supervisors, senior managers, etc.) to engage in knowledge-sharing behavior; on the other, they underline the importance of enhancing the individual's sense of self-esteem.

Studies concentrating on the patterns that characterize annotation efforts in Web 2.0 communities found an interesting correlation between a set of emerging social roles and tagging behaviors. These behaviors seem to be spurred by the attempt to create a community, the awareness of one's audience and a perceived need to communicate with a small group [22]. Analogously, Chen and colleagues found that social comparisons help explain the tendency to contribute more (or less) in a social experiment involving the MovieLens community [6]. Joinson identified these unique uses and gratifications in the context of Facebook: social connection, shared identities, content, social investigation, social network surfing and status updating [13].

From these examples, we can clearly see that motivation can be produced by heterogeneous motives, and might derive from incentives that are assigned to the performer or from an intrinsic desire. Motivation is intrinsic if the performer enjoys the act of performing the task *per se*. In all other cases, a set of extrinsic incentives can be provided in order to make an individual/team perform. Incentives are a set of instruments (e.g., money, reputation, rewards, prices, credit points, medals) assigned by an external 'judge' typically according to some sort of evaluation of the effort exercised by the performer. In principle, these can be totally uncorrelated to the nature of the task.

B. Mechanism design theory

Mechanism design is a field of game theory developed in economics that studies the effective design of rules for human behavior. If individuals follow these rules, they will reach the outcome desired by the game designer. The

underlying hypothesis is that individuals act according to their own private interests and only a careful development of appropriate incentives can enable the alignment of individual and social interests.

To develop a set of incentives from the mechanism design point of view, the first thing developers should do is to understand the social environment (the context) and codify its constraints from the point of view of game theory. In these terms, the game is defined by the following features [19]: the players, the rules, the outcomes and payoffs. In the case of Web 2.0 applications the players are the users (both contributors and readers of the content produced), the rules refer to how the players interact among each other, the outcomes are constituted by the public good produced by means of the application, and finally the payoffs are the values players attribute to the outcomes.

In the analysis of the context, the designer should focus on the system of individual inner interests and the motivations embedded in the structure they interact with (the tool for instance). These interests are affected by various elements which are:

- *The goal:* people interact to communicate and participate.
- *The nature of good produced:* a stylized description, in game-theoretical terms, of the relationship between what good is produced and who consumes it.
- *The tasks:* an ordered collection of tasks into which the contributions can be broken down.
- *The skills:* competences and abilities required to carry on the tasks.
- *The social structure:* a stylized and simplified set of social relationships among the subjects participating in the exercise.

After the analysis of the context it is necessary to define the desirable outcome that the designers want to achieve. Based on the given definition of players, desirable outcome, and the context it is possible to define a set of rules and payoffs that permit to achieve the desirable outcome.

Looking at Web 2.0 applications users spend time and effort to produce - and at the same time to consume - a public good. All individuals benefit from the outcomes that others produce and the application provides. The possibility to access the content without necessarily contributing to its creation leads to the phenomenon of free riding: expecting that others will spend their time and effort to create the content while dedicating own time to other rewarding activities. Mechanism design tools allow us to analyze the context corresponding to each Web 2.0 application and to design some case-specific rules that lead to a reduction of this kind of behavior within the desired context.

III. HOW TO DESIGN AN INCENTIVIZED APPLICATION

Mechanism design enables to analyze the social structure of the scenario prior to application design, to fine tune the

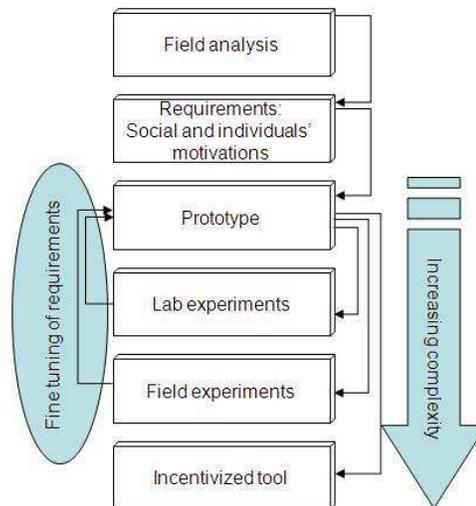


Figure 1. The ideal process of design and development of an incentivized application

incentive structure of an application, and to provide a set of requirements for a first prototype of the incentivized tool. The lab experiments, then, allow us to evaluate such a tool and to fine tune the incentive structure that is embedded into it.

Ideally the process of design and development of an incentivized application (see Figure 1) should start from the analysis of the concrete situation.

In the first phase, the field analysis is crucial to identify the motivations of both individuals and the social groups which they belong to. Direct observations, interviews and questionnaires are very effective techniques that can be used to unveil and better define the crucial elements described in the previous section.

In the second phase, the raw knowledge is then analyzed in terms of the above mentioned elements. Mechanism design, as a set of techniques, allows the modeling of the situation by using game theoretical predictions about the behaviors of the actors described in the model. Given a set of goals, this model enables the analysts to design a set of incentive schemes that would spur users to behave in line with the desired outcomes.

The third phase is the creation of the prototype which should be the simplest possible solution that can effectively support the users.

In the fourth phase, the resulting prototype is tested, better if the test is conducted in a controlled environment, such as a laboratory. The laboratory allows the experimenter to test an hypothesis with artificially controlled conditions, manipulating or eliminating extraneous factors. As soon as the previous hypotheses are confirmed a sequence of experiments can be organized to fine tune the set of incentives that are embedded in the tool. The design of each experiment may depend on the results of previous ones.

In the fifth phase, the field experiments, the tool is tested in the field, mimicking the situation in the lab. This fine tuning process should increase the complexity of the trial since the experiment get closer to the reality. For instance, add new realistic components such as real actors (i.e., community members), tasks (a daily activity that actors usually carry on), situations (the field and the social structure which actors belong to). Since the control over the ability to manipulate variables reduces, the experiments gain awareness of interaction among variables. This might continuously address new changes to the tool that finally is introduced in the field.

This process is continuously repeated since the tool is finalized, sixth phase.

IV. THE ANALYSIS OF AN ANNOTATION PROTOTYPE DEVELOPED BY T+ID

Founded in 1988, Telefónica Investigación y Desarrollo (TID) [16] is nowadays the largest private research and development centre in Spain as regards activity and resources, and is the most active company in Europe in terms of European research projects in the Information and Communication Technology sector. TID is currently developing a prototype for the annotation of the internal portal of Telefónica. The tremendous amount of available information hinders the access to the right pieces of information required by a person for a specific goal, affecting the company workflows. This is why the company decided to develop an annotation tool to the corporate portal. Adding a layer of annotations, helps obtain many advantages, such as more efficient resource retrieval and navigation, real integration of heterogeneous sources of information (linked data), personalization based on context and roles.

In the following section, we will go through the steps we adopted to support the design of TID's annotation tool. It will serve as an illustration of the general process we described in the previous section.

A. *The first phase: the field analysis*

We started our analysis collecting data on people's motives and motivation drivers interviewing 11 employees of TID representing the community at large (heads of division, senior project managers, project managers, developers, computer engineers, and consultants). Each semi-structured interview was conducted by two interviewers, took 60 to 90 minutes and was recorded on audio tape. These recordings have been transcribed and analyzed descriptively according to ex-post categories. Additionally, a focus group discussion with 6 TID employees was conducted, focusing on usage problems of the existing system and on innovative solutions that might overcome these problems. Since the number of interviewees was not very high, the TID interviews were decisive to provide starting data for the desk analysis phase.

More in depth requirements are identified in the following phases.

B. *The second phase: the desk analysis*

We analyzed the interviews and classified TID portal's features according to the variables described above.

The goal of the annotation tool is to improve the search and navigation experience in a corporate knowledge base. TID providers of annotations might have two different motivations to annotate. Users - by annotating resources - can show their areas of competence and interest to the community. They improve navigation, searching and syndicating capabilities of the enterprise portal by using annotations.

The nature of good produced by annotating is public. This scenario is an almost straight out of the textbook case of public good provision in which providers and consumers are the same people. The part where things become problematic is the problem shared by many knowledge management systems: there is a huge incentive to keep strategic knowledge private so one can leverage on it when dealing/negotiating with others.

The task is a typical annotation task. This means that it is very repetitive and lacks a fun element.

The required skills of the agents to complete the annotation task are very basic.

Finally, the social structure is quite complex and various dynamics coexist. Employees work in teams and communities of experts, but also work in the company with a strong hierarchical structure. Visibility, reputation, career development, and money are all part of the mix of motivations driving the behavior of employees. An important issue to deal with is the large number of employees. As underlined in many studies, group size plays an important role in modifying behavior of individual contributors. If, in principle, employees perceive that their contribution is vital for the success of the group we could expect a higher probability of contribution from each employee. In other words, the reputation mechanism might be developed at group or project level.

In this paper, we do not focus our analysis on the tool development, therefore we do not discuss the third phase: the prototype development process.

C. *The fourth phase: the lab experiments*

In the evaluation of the first prototype we run a laboratory experiment with the goal of identifying the more appropriate set of incentives to spur people to annotate resources.

The incentive structures we tested are focusing on two different rewarding systems:

- pay per click models: each participant get a fixed amount of money for each annotation provided (0,03 Euros per tag up to 3 Euro). For payment requirements we had to round rewards in pay per tag treatment to the highest 50 cents.

- winner takes all model: only the actor that provide the higher number of tags and annotation wins 20 Euros.

Participants were 36 students with no experience in the annotation task nor in the tool that they tested. At the beginning participants read a clear set of instructions and performed a first training session in order to understand the rules of the game and the basic features of the annotation tool. Each individual has been assigned to one of two 'treatments' of images annotation: the first using the pay per click model and the second using the winner takes all models. They performed the task under time pressure (8 minutes for each treatment) with the goal to produce the maximum amount of tags in the allotted time on a random set of images.

We obtain the following results:

- The mean number of tags produced are: 47.42 with the pay per tag model and 62.76 with the winner takes all model. We can measure a 32% increase in the winner takes all model.
- The average amount of money participants perceived are: 6.66 Euros with the pay per tag model and 6.18 Euros with the winner takes all model.
- The average cost per tag is 0.1404 Euros with the pay per tag model and 0.098407 Euros with the winner takes all model.
- The budget has been 31.5 Euro for the pay per tag model and 20 Euros for the winner takes all model.

In the experiment there are some biases. Students are volunteers who are used to participate in experiments. Also, students didn't find any problem in the annotation task because they are strong Web users and game players.

D. The fifth phase: the field experiments

Despite the biases that affect the experiment results, it clearly emerged that the winner takes all model is dramatically more effective than the pay per click one, in this context. These results constitute a baseline for the implementation of the tool within TID's corporate environment. The obvious next step is that of running a few other lab experiments in order to move from simple tagging to more complex tasks closely mirroring what happens in TID.

Since annotation is used to retrieve, organize and exchange information in the company portal, in the next experiments annotations and the resulting tags obtained in a first stage of the experiment will be used as an input to retrieve information in the fifth phase: the field experiments.

In the TID, the prototype and the set of incentives will be tested first with a small group of users in order to fine tune the adequate set of requirements. Eventually, the finalized version of the incentivized application will be applied to the whole TID.

V. CONCLUSION

As mentioned above, one of the dominant traits of the Web 2.0 applications is the capability of co-opting end-users

in creating new content and sharing it across their networks. The tremendous growth of data, pictures, images, videos that are shared and copied by individuals in the networks, render the access to the right piece of information more difficult. As a consequence many Web 2.0 applications, such as Facebook and others, introduce annotation tools enabling users to add, modify or remove information about a Web resource without modifying the resource itself.

In this paper we demonstrate that, since the whole software development process is very relevant, the success of social software tools requires carefully crafted social design policies aimed at fostering participation and willingness to contribute.

For this purpose, mechanism design and laboratory/field experiments seem promising methods to enable designers to analyze the motivations of users and embed this information into their software.

These phases should be seen as a continuous improvement process that enables designers to adjust the software according to the social needs emerging from the users' experiences.

As we have just seen, we have proposed a methodology that encompasses the ability to a) analyze any work environment of a social nature in which one aims to introduce annotating tasks; b) conduct a design process for the tools that serve as vehicle for the tasks themselves taking the previous analysis of the social and technological context as an input. This approach stresses a value-chain outlook on the design process, clearly distinguishing problems regarding motivation/participation in the design process *strictu sensu*, on the one hand, and motivation/participation of users once the tool is in place. The most glaring advantage of this approach is the ability to consolidate theory, method and applications in distinct units, avoiding redundancy.

Finally, since the methodology we propose seems promising and accepted by developers, the real benefits of the incentivized application will be measured only when it will be adopted by TID.

ACKNOWLEDGMENT

This work was supported by the EU funded project IN-SEMTIVES - Incentives for Semantics (www.insemtives.eu, FP7-ICT-2007-3, Contract Number 231181).

REFERENCES

- [1] Adar, E. and Huberman, B. A., Free Riding on Gnutella, *First Monday*, October 2000. Available at: <http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/-792/701>.
- [2] Beenen, G., Ling, K., Wang, X., Chang, K., Frankowski, D., Resnick, P., and Kraut, R. E., *Using Social Psychology to Motivate Contributions to Online Communities*, Proceedings of the CSCW'04, November 6-10, Chicago, Illinois, USA, pp. 212-221, 2004.

- [3] Bock, G. W., Zmud, R. W., Kim, Y. G., and Lee, J. N., 'Behavioral Intention Formation in Knowledge Sharing: Examining the Roles of Extrinsic Motivators, Social-Psychological Forces, and Organizational Climate', *MIS Quarterly*, 29 (1), pp. 87-111, 2005.
- [4] Bouman, W., de Bruin, B., Hoogenboom, T., Huzing, A., Jansen, R., and Schoondorp, M. *The Realm of Sociality: Notes on the Design of Social Software*, Proceedings of the International Conference on Information Systems (ICIS), Montreal, Canada, paper 154, 2007.
- [5] Cattuto, C., Loreto, V., and Pietronero, L. *Semiotic Dynamics and Collaborative Tagging*, Proceedings of the National Academy of Sciences of the United States of America, 104, pp. 1461-1464, 2007.
- [6] Chen, Y., Harper, F. M., Konstan, J., and Xin Li S. *Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens*, Working paper, School of Information, University of Michigan, pp. 1358-1398, 2008.
- [7] Davenport, T. H. and Prusak L. *Working Knowledge: How Organizations Manage What They Know*, Cambridge, MA: Harvard Business School Press, 1998.
- [8] Fang, Y. and Neufeld, D. 'Understanding Sustained Participation in Open Source Software Projects', *Journal of Management Information Systems*. Spring, (25)4, pp. 9-50, 2009.
- [9] Forte, A. and Bruckman, A. *Why do people write for wikipedia? Incentives to contribute to open-content publishing*, in Proceedings of 41st Annual Hawaii International Conference on System Sciences (HICSS), pp.119-128 2008.
- [10] Golder, S. A. and Huberman, B. A. 'The Structure of Collaborative Tagging Systems', *Journal of Information Science*, 32, pp. 198-208, 2006.
- [11] Hars, A. and Qu, S. 'Working for free - Motivations for participating in open-source projects', *International Journal of Electronic Commerce*, 6, pp. 25-39, 2002.
- [12] Howe, J. *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*, Crown Publishing Group, New York, NY, 2008
- [13] Joinson, A. N. *Looking at, looking up or keeping up with people? Motives and uses of Facebook*, Proceedings of 26th Annual ACM Conference on Human Factors in Computing Systems, pp. 1027-1036, 2008.
- [14] Kleeman, F., Voss, G. G. and Rieder, K. Un(der)paid Innovators: The Commercial Utilization of Consumer Work through Crowdsourcing. *Science, Technology and Innovation Studies*, (4) 1, pp. 5-26, 2008.
- [15] Kuznetsov, S. *Motivations of contributors to Wikipedia*, ACM SIGCAS Computers and Society, (36) 2, Article 1, 2006.
- [16] Lopez, Garcia, S., The Role of Telefonica: The Internationalization of Telecommunications in Spain, 1970-2000, *Business and Economic History*, (1), 2003. Available at: <https://35.9.18.4/business/bhweb/publications/BEH-online/2003/Lopez.pdf>.
- [17] Malone, T. W., Laubacher, R. and Dellarocas, C. N. *Harnessing Crowds: Mapping the Genome of Collective Intelligence*, MIT Sloan Research Paper No. 4732-09, 2009. Available at: <http://ssrn.com/abstract=1381502>.
- [18] O'Reilly T. *What Is Web 2.0 Design Patterns and Business Models for the Next Generation of Software*, 2005. Available at: <http://oreilly.com/web2/archive/what-is-web-20.html>.
- [19] Osborne, M. J. and Rubinstein, A. *A course in Game theory*, The MIT Press, 1994
- [20] Smith, D. M. *Key Issues for Web 2.0 and Consumerization*, Gartner Research Report, 2007. Available at: <http://www.gartner.com>
- [21] Sugden, R. 'Consistent Conjectures and Voluntary Contributions to Public Goods: Why the Conventional Theory does not Work', *Journal of Public Economics*, (27), pp. 117-124, 1985.
- [22] Thom-Santelli, J., Muller, M. J., and Millen, D. R. *Social tagging roles: Publishers, Evangelists, Leaders*, Proceedings of 26th Annual ACM Conference on Human Factors in Computing Systems, pp. 1041-1044, 2008.
- [23] Von Hippel, E. *Horizontal Innovation Networks - by and for users*, *Industrial and Corporate Change*, (16) 2, pp. 293-315, 2002.
- [24] Wang, Y. and Fesenmaier, D.R. 'Assessing motivation of contribution in online communities: An empirical investigation of an online travel community', *Electronic Markets*, 13, pp. 33-45, 2003.
- [25] Wagner, C. and Prasarnphanich, P. *Innovating collaborative content creation: The role of altruism and wiki technology*, Proceedings of the 40th Annual Hawaii International Conference on System Sciences, pp. 18-27, 2007.
- [26] Wasko, M. and Faraj, S. 'It is what one does: Why people participate and help others in electronic communities of practice', *Journal of Strategic Information Systems*, (9) 2-3, pp. 155-173, 2000.
- [27] Wasko, M. and Faraj, S. 'Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice', *MIS Quarterly*, 29 (1), pp. 35-57, 2005.
- [28] White, D. *Results and analysis of Web 2.0 services survey*, UK:JISC, 2007. Available at: <http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/spiresurvey.pdf>
- [29] Wiertz, C. and de Ruyter, K. 'Beyond the Call of Duty: Why Customers Contribute to Firm-Hosted Commercial Online Communities', *Organization Studies*, 28 (3), pp. 347-376, 2007.

Performance Evaluation of Dynamic Web Service Selection Strategies in Service Oriented Architecture

Miroslav Živković *, Hans van den Berg, Hendrik B. Meeuwissen, Bart M. M. Gijssen

TNO

Delft, The Netherlands

miroslav.zivkovic@tno.nl

* also with

Design and Analysis of Communication Systems

University of Twente

Enschede, The Netherlands

m.zivkovic@utwente.nl

Abstract—In this paper, we explore the performance potential of dynamic (runtime) web service selection within the scope of Service Oriented Architecture (SOA). The web service selection is executed by a service orchestrator (request dispatcher) which is responsible to deliver composite web services at desired quality levels for the orchestrator’s clients. We investigate service response times for the case where SOA state-of-the-art static web service composition is used and for the case where dynamic web service selection is applied. Modelling request scheduling at individual web services as Processor Sharing queueing systems, simulation results are presented for different runtime selection strategies in scenarios ranging from the “ideal” situation (up-to-date state information, no background traffic) to more realistic scenarios in which state information is stale and/or background traffic is present. In particular, we show the effectiveness of a selection strategy based upon the “synthesis” of Join the Shortest Queue and Round Robin strategies. For some specific scenarios we derive and validate insightful approximate formulae for the resulting response times. Our investigations quantify the performance gains that can be achieved by dynamic service selection compared to static (a-priori) service selection currently used.

Keywords—Service Oriented Architecture, Join the Shortest Queue, Processor Sharing, Response Time, Background Traffic, Stale Information.

I. INTRODUCTION

The composition of web services within a SOA environment could be static or dynamic. With static composition the concrete services are determined and integrated into the specification at design time. With dynamic composition on the other hand, at design-time there is only a specification of the required abstract services given. The concrete services are then integrated at runtime.

For both static and dynamic composition, the choice of concrete services for a particular abstract service may be based on non-functional parameters. Examples of such parameters are availability, throughput, response time, security and cost. References [4], [21], [23] and [24] discuss the problem of static QoS-aware service composition in detail.

As an extension, [4] considers sub-optimal algorithms to enable fast replacement of underperforming services.

Existing SOA-based solutions for web services do not support dynamic web service selection ([12], [13]). Dynamic selection provides flexibility and therefore has advantages with respect to availability and reliability compared to a static approach. Another possible advantage of dynamic service selection is performance improvement, i.e., achieving a decrease in the requests’ response times by exploiting statistical variations in the loads at the various concrete services. This paper aims at investigating the potential performance gain of dynamic, runtime web service selection for service composition within the scope of SOA, evaluating different selection strategies.

As a starting point, we assume that a *set* of concrete services is a-priori selected per abstract service, which is in contrast with a-priori selecting a *single* concrete service per abstract service in the case of static composition. Therefore, the discovery mechanisms ([19]) and their performance are outside the scope of this paper. As an example, in Figure 1, the choice of a particular concrete service (from the set of selected concrete services) is made by the dispatcher at runtime on a per-client request basis for a composite web service consisting of N services that are invoked consecutively. In each consecutive step i , $i = 1, 2, \dots, N$, exactly 1 out of K_i concrete services is invoked by the dispatcher, where K_i represents the number of choices in step i . In this example, the total response time RT_{total} is the sum of the individual response times RT_i . Within SOA the dispatcher is part of the orchestrator, which typically runs in the domain of the composite service provider.

The main research question addressed in this paper is what is the performance potential of dynamic web service selection versus static selection? As outlined above, dynamic web service selection is made from K_i pre-selected concrete services for each step i (note that this pre-selection is beyond the scope of this paper). We focus on the achievable performance gain of dynamic selection in case of a single

abstract service (i.e., $N = 1$ in Figure 1). This analysis will give us also an indication of the potential gain for a composite service that consists of more abstract services ($N > 1$). Besides, analysis for the case $N = 1$ is more feasible than for the general case $N > 1$. Specific research questions addressed in the paper are: what is the influence of the number of pre-selected concrete services, which stateless and statefull dispatching algorithms perform well, which gains are achievable, and what is the impact of practical conditions such as background traffic and delayed state information on these gains?

Notice that existing dispatching strategies, e.g., Round-Robin (RR), Bernoulli, Join the Shorted Queue (JSQ), are included in our analysis. In the literature, these dispatching strategies are mostly investigated in the context of systems with First Come First Served (FCFS) queues, but hardly for systems with Processor Sharing (PS) queues, as considered in this paper. Note that in the current context the PS services model is more realistic than FCFS, see e.g., [10]. In addition, we consider background traffic and delayed state information.

We summarize the main contributions of this paper as follows:

- 1) Quantification of the achievable gain versus the number K of pre-selected concrete services by a fair comparison with respect to the base case of static web service selection.
- 2) Quantification of the achievable gain in terms of response times when background traffic is present at the pre-selected concrete services for different dispatching strategies.
- 3) Quantification of the achievable gain in terms of response times in case of delayed state information.

In order to investigate the above-mentioned potential performance gains several assumptions have been made, which allow us to quantify the “ideal” system performance. In that sense our analysis should be understood as a “baseline” analysis. Besides, the assumptions are made with the goal to represent our findings in an unequivocal way. Our analysis makes a significant step towards analysis of models that take into account more practical conditions regarding the observed system.

The remainder of this paper is organized as follows. First, in Section II, we describe the performance model and explain the underlying assumptions that capture the essential system characteristics needed for our study. Next, in Section III, we discuss literature related to our work. In Section IV, our simulation results and results obtained by analytical modelling are presented, and we discuss and explain the observations. Finally, in Section V, we draw conclusions and give suggestions for further research.

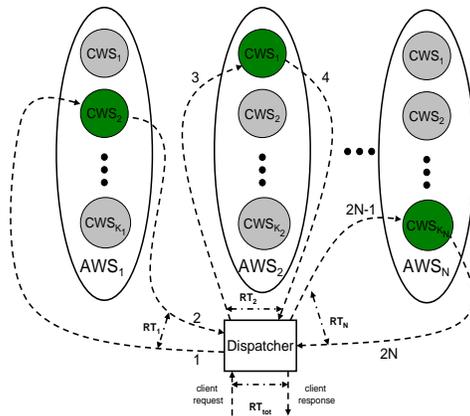


Figure 1. Illustration of dynamic SOA-based web service composition. A request by the client is served by one of the K_1 implementations of abstract service AWS_1 , then by one of the K_2 implementations of abstract service AWS_2 , and so on, until client request is completed and response is sent back. Note that every client request may be served by a different chain of concrete web services (CWS).

II. MODEL DESCRIPTION

We consider one abstract service with K concrete service implementations as given in Figure 2. There are two classes of incoming service requests:

- Foreground service requests are received by the dispatcher, which decides at runtime to which of the K service instances a particular request is assigned for getting the required service. The foreground requests arrive to the dispatcher according to a Poisson process [15] with rate Λ and have exponentially distributed service requirements with mean $\frac{1}{\mu}$. The rate at which foreground traffic requests are offered (by the dispatcher) to service i is denoted by λ_{FTi} , $i = 1, 2, \dots, K$.
- Background service requests arrive at service instance i according to a Poisson process with rate λ_i , $i = 1, 2, \dots, K$, respectively. The background service requests are exponentially distributed with mean $\frac{1}{\mu_i}$, $i = 1, \dots, K$. The background traffic arrival processes are independent from each other and are also independent from the foreground arrival process.

Request scheduling at each service instance is modelled by a processor sharing infinite-capacity single server queue. Served requests leave the system.

Obviously, the achievable performance gain of dynamic service selection compared to a pure static approach depends heavily on the nature of the workload fluctuations at the different service instances. It is clear that in the case of slowly and independently varying loads (e.g., due to fluctuations in the service demand over the day) high performance gain can relatively easy be achieved. However, in such cases the runtime character of dynamic service selection may be

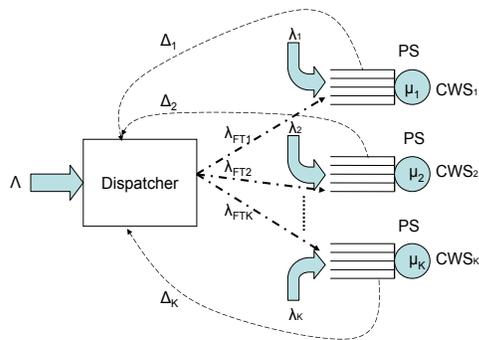


Figure 2. Performance model for the case of a single abstract service with K different implementations (concrete services).

an “overkill” and the performance gain could largely also be obtained by less flexible service selection approaches. Therefore, the present paper focuses on exploiting workload fluctuations at the services instances that occur at relatively small time scales mainly caused by the random behaviour of individuals in a large population of potential users. In that perspective, and to keep the parameter space manageable, we will assume that the model is symmetric, i.e., $\lambda_i = \lambda_{BT}$, $\mu_i = \mu$, $i = 1, 2, \dots, K$. The utilization per service i , $i = 1, 2, \dots, K$ is then defined as $\rho_{tot} = \frac{\lambda_{tot}}{\mu}$, where λ_{tot} is the rate of the aggregated (foreground and background) traffic, i.e., $\lambda_{tot} = \lambda_{FT1} + \lambda_i = \lambda + \lambda_{BT}$. The stability condition per service requires that the expected number of requests per service remains finite, i.e., $\rho_{tot} < 1$.

Ignoring possible delays due to the queueing and processing at the dispatcher, as well as network delay, arriving service requests are instantaneously forwarded to one of the K service instances according to the dispatching strategy. Various strategies can be used for selection of one of the K service instances upon arrival of a new request. The dispatching strategies could be roughly divided into two categories, namely *stateless* and *statefull*. Decision making is independent of the system state information for the stateless strategies. Conversely, decision making takes into account (stale) system state information for statefull strategies. The delay in obtaining the information per service instance is represented by parameter Δ_i , $i = 1, 2, \dots, K$. In this paper we have adopted the case when system state information (queue length, response time, etc.) is collected periodically with the same period $\Delta > 0$. This information gathering may require sending separate requests (“probes”) by the dispatcher to all of the K web services, and collecting information in such a manner introduces an overhead to the system, which influences system performance. This issue as well as using other ways to collect system state information are beyond the scope of this paper. The update period Δ has been related to the intensity of the aggregated traffic as $\Delta = D \cdot \frac{1}{\lambda_{tot}}$, where D is integer. All dispatching decisions

between time instances $t_i = i \cdot \Delta$ and $t_{i+1} = (i+1) \cdot \Delta$, $i = 0, 1, 2, \dots$ are made based on the system state information obtained at t_i . The dispatching strategies considered for this paper are Bernoulli (BL), Round-Robin (RR), Join the Shortest Queue (JSQ) and a combination of the latter two, JSQ–RR. Bernoulli and RR are typical examples of stateless strategies while JSQ and JSQ–RR are examples of statefull dispatching strategies. In case of the Bernoulli strategy, the requests are randomly distributed over the queues, i.e., a newly arriving request is assigned to queue i , $i = 1, \dots, K$, with probability $\frac{1}{K}$. This case is used as representation of the performance for the static SOA service selection. For RR, the k -th request is assigned to queue $(k \bmod K) + 1$. In JSQ, the request is assigned to the queue with the smallest number of requests waiting to be served. Ties are resolved by randomly assigning the request to one of the shortest queues. In case when system state information delay is present in the system, an additional statefull dispatching strategy could be defined, namely JSQ–RR. For JSQ–RR, once the actual system information is obtained, the queues are sorted in non-descending order by the queue lengths. Any request coming to the dispatcher between two state updates is then assigned following the RR scheme, i.e., the first request is assigned to the queue with smallest queue length, the second request is assigned to the queue with smallest queue length from the remaining queues, etc.

III. RELATED WORK

In this section, we give a short overview of papers related to different aspects (e.g., web service selection, composition, performance) of the runtime web service selection in SOA. However, each of these papers treats only a (different) subset of issues relating to runtime web service selection. The analysis of potential performance improvements of different dispatching strategies, based on PS modelling of the request scheduling at web service(s) within SOA, which includes impact of stale system state information and/or background traffic is, to our best knowledge, non-existent.

A. Web service selection and composition

In [13], an overview of common misconceptions about SOA is given. Among others, the issue of dynamic selection of web services is identified, and it is indicated that current SOA solutions lack advanced automatic discovery and composition of web services at runtime.

A lot of attention within SOA community has been dedicated to static QoS-aware composition problem, see e.g., [4], [21], [23] and references therein. The problem of static QoS-aware composition is known to be NP-hard, see [24], where two service selection approaches for constructing composite services have been proposed: local optimization and global planning.

In paper [20], several architectures and their respective models that assist in dynamic invocation of web services

are discussed. These models allow the *client* to dynamically select the current best web service, based on certain non-functional criteria (availability, reliability, and estimated response time). These clients gather runtime web service information, evaluate the performance of the previously used web services, and may share this information with other clients. The selection decision is let to the clients, which contain intelligent agents and therefore the complexity of the clients increases. The inherent problem is that different clients may decide to use the same web service, which would eventually result in worsened performance e.g., due to the overload of the targeted web service.

The framework proposed in [14] enables quality-driven web service selection, based upon evaluation of the QoS of a vast number of web services. The fair computation and enforcing of QoS of web services takes place when making the web service selection. In order to provide fair computation the feedback from clients is gathered.

B. Performance of dispatching strategies

Performance of dispatching strategies in multi-server systems has been a topic that received a lot of attention within the queueing theory research community. Specifically, a lot of work has been done for systems with First Come First Served (FCFS) scheduling at the queues, e.g., in [3] and [6]. In the most of the papers written for JSQ/FCFS, explicit results for response times are given only for the case $K = 2$ servers, an exponential job size distribution and the mean response time metric, [9]. The performance of the JSQ/FCFS strategy for $K > 2$ servers has been analysed in [17] where the approximation of the mean response time for K homogeneous servers is given. In [18], an extension to this approximation has been given, however, the approximation is less accurate as the requests' size variability increases.

Opposite to the JSQ/FCFS systems, JSQ/PS systems have not received so much attention. The notable exceptions are [2], and, more recently, [10] where approximate analysis of JSQ in the PS server farm model for general job size distributions is presented. The queue length of each queue in the system is approximated by a one-dimensional Markov chain, and based on this approximation the distribution of the queue length at each queue is determined. In [1], the authors investigate optimal dispatching strategies for a multi-class multi-server PS systems with a Poisson input stream, heterogeneous service rates, and a server-dependent holding cost per unit time.

C. Performance of dispatching strategies with stale system information or background traffic

In [16], the problem of dispatching with stale system status information (server load) is analysed in case of FCFS. Servers' status information is periodically updated and three strategies are compared: random selection, selection of the

server with the least load (based on the stale system information), and random selection of a small subset of servers and then selecting the least loaded of the chosen servers (based on up-to-date information about their loads). It is shown that the latter strategy mostly outperforms the other ones, even for a small randomly chosen subset of e.g., two servers, while the overhead (due to processing and information retrieval) remains limited. In [5], the authors present a strategy that routes the jobs to the server with expected shortest FCFS queue. The decisions are made based on stale information and elapsed time since the last state update. This strategy works well, but does not always minimise the average response time.

In [11], a dispatching policy based on splitting foreground traffic according to a predefined rule described by a certain parameter vector is analysed while background traffic is modelled as independent Poisson processes with different rates. Due to the assumptions made each of the N servers in isolation can be represented as a two-class M/G/1 PS queue. The approximation of the response times is deduced for the case of light foreground traffic and an optimal parameter vector is found.

IV. PERFORMANCE ANALYSIS

In this section, we present and discuss simulation results for the runtime service selection strategies described in Section II in order to investigate their performance potential. For some special scenarios we also present numerical results obtained from analytical modelling.

The simulations were performed using the simulation tool implemented in Java programming language, and using the Java library for stochastic simulation (SSJ) [8]. In order to make the simulations less sensitive to the startup transient, the number of foreground traffic arrivals per simulation has been set to at least $0.5 \cdot 10^6$. Besides, in order to improve the accuracy, we have trimmed simulation results for certain number of foreground traffic arrivals at the end of the arrival process.

We have considered four main categories of simulation scenarios:

- Baseline scenarios – these simulations were performed for the system without background traffic and with up-to-date system state information. The simulation results are given in subsection IV-A.
- Scenarios with stale system state information – these simulations were performed for the system without background traffic in which system state information is only periodically updated i.e., the dispatching process does not (always) use up-to-date information. The results are presented in subsection IV-B.
- Scenarios with background traffic – these simulations were performed for the system with up-to-date system state information and different intensities of background traffic. In addition to the simulations we derived

an analytical approach to study the performance for these scenarios. The results are presented in subsection IV-C.

- Scenarios with background traffic and stale system state information. The simulation results are presented in subsection IV-D.

A. Baseline scenarios

The goal of these simulation scenarios was to establish the performance results in case when there is no background traffic and system state information is up-to-date at the dispatcher.

In Figure 3, we show mean response times for different dispatching strategies (JSQ, RR, BL) as a function of the number of concrete services, K and for different values of utilization per service, ρ_{tot} . The utilization per service is kept constant in order to have a fair assessment of the impact of K ; otherwise, an increase of K would simply be interpreted as capacity add-on to the system. Since JSQ-RR is identical to JSQ when up-to-date system state information is available at the dispatcher, the results for JSQ-RR are not shown. For $\rho_{tot} = 0.8$, the mean response time for JSQ strategy with $K = 4$ services is around 66% of mean response time for the same strategy when $K = 2$. Similarly, in case of JSQ with $K = 8$ and $K = 16$ services, response times are 49% and 40% of the response time for $K = 2$, respectively. In case when one of the K services becomes unavailable, the performance of the system (response time) does not deteriorate dramatically, as long as the utilization per queue remains (approximately) the same. The utilization per queue can be kept the same when, e.g., $K + 1$ services are pre-selected, of which given (fixed choice) K services are used for dispatching. The remaining ($K + 1$ th) service is placed "on hold" and when one of the chosen K services becomes unavailable, it is immediately replaced.

Figure 4 shows relative comparisons between JSQ and BL (with BL as the baseline) and JSQ and RR strategies (with RR as the baseline), respectively. Statefull strategy (JSQ) is superior to either of the stateless strategies (BL, RR), which confirms that more (and accurate) state information made available to the dispatcher leads to better decision making.

The potential performance improvements in the first case range from 28% to 46% for $K = 2$, depending upon the utilization per queue, ρ_{tot} and are in the range from 49% to 86% for $K = 16$. What is also of interest is when do the gradient of the performance improvement is highest, taking into account the increase of the number of services. From Figure 4 we see that this is the case when the number of services increases from 2 to 4. The gradient of the gains is (significantly) smaller when the number of services increases from 4 to 8 or 8 to 16, respectively. Based on these simulation results, we can draw the following conclusions:

- Large performance improvements compared to the static service selection are possible with relatively small

values of K .

- The largest relative improvements of response time for number of services, $K > 1$, are obtained when we increase the number of services that are used from 2 to 4.

B. Scenarios with stale system state information

The goal of these simulations was to analyse the impact of the stale system state information to the response time of the system for different dispatching strategies. No background traffic has been assumed. The simulations were performed only for the statefull strategies, i.e., JSQ (see Figure 5) and JSQ-RR, see Figure 6. The response times for stateless strategies, RR and BL, are not affected by (stale) system state information, and are shown for comparison as well.

From Figure 5 we see that, for relatively small values of parameter D that determines the update interval, JSQ still performs better than RR or BL dispatching strategy. However, as expected, when parameter D increases performance of JSQ deteriorates e.g., for $D = 20$ response time for JSQ is worse than either RR or BL for almost complete range of parameter ρ_{tot} . In case when $D \rightarrow \infty$, the system state information is obtained just once, and then all arrivals are "blindly" assigned to the queue which had the smallest queue length when system state information was obtained. In that case, the service composition in this case is static, and the system model reduces to a M/M/1/PS queue with arrival rate Λ and mean service time μ .

We have also investigated the behaviour of the JSQ-RR strategy for systems with stale state information. Figure 6 shows that, as expected, JSQ-RR strategy is less sensitive to stale information than "blind" JSQ strategy. For example, when $D = 10$ and $\rho_{tot} = 0.7$, response times for Bernoulli, RR, JSQ and JSQ-RR are 2250 ms, 1515 ms, 3200 ms (!) and 1350 ms, respectively. For comparison, the response time for JSQ without stale information and the same ρ_{tot} is approximately 1020 ms.

Based on the simulations which results are presented at Figure 5 and Figure 6 we can draw the following conclusions:

- When $D \rightarrow 0$ JSQ-RR is identical to JSQ and when $D \rightarrow \infty$, JSQ-RR is identical to the common RR strategy.
- With respect to the response time, the JSQ-RR strategy is never worse than RR, regardless of the delay within the system. This makes JSQ-RR appealing strategy for systems with delay without background traffic.

C. Scenarios with background traffic

In the previous simulation scenarios we have assumed that the concrete services were used by the foreground traffic clients only. In what follows we look into the situation when background traffic is present as well, and the dispatcher has up-to-date system state information.

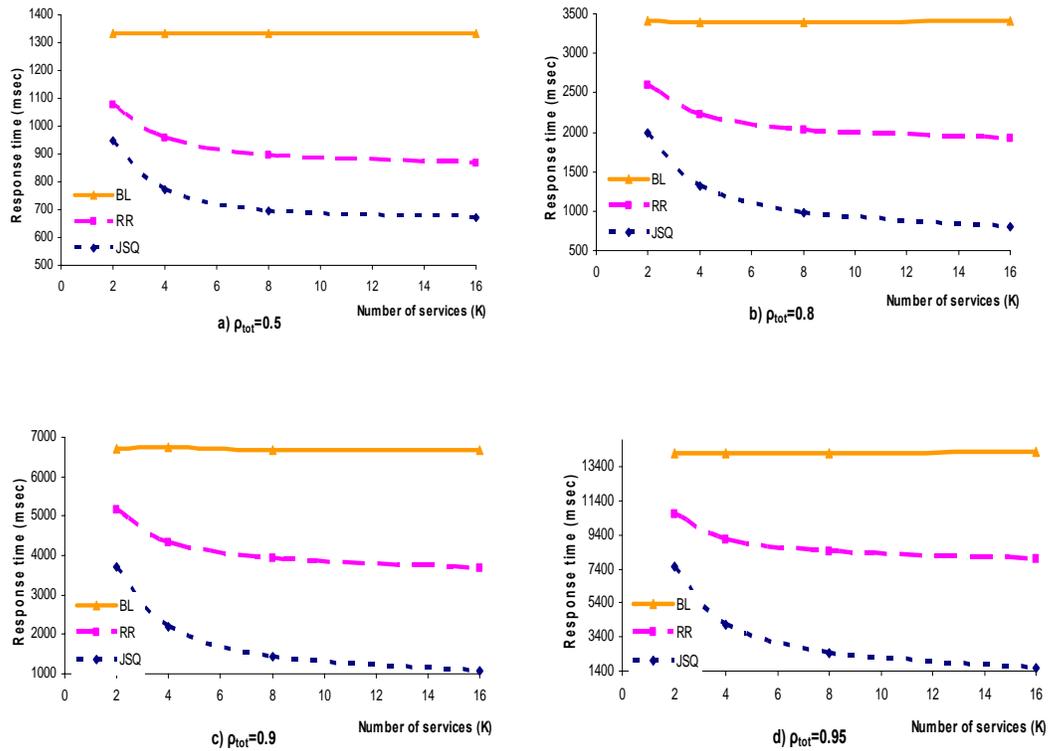


Figure 3. Comparison of mean response times for JSQ, RR and BL strategies for different number of services K and different values of ρ_{tot} . There is no system state information delay and only foreground traffic is present in the system.

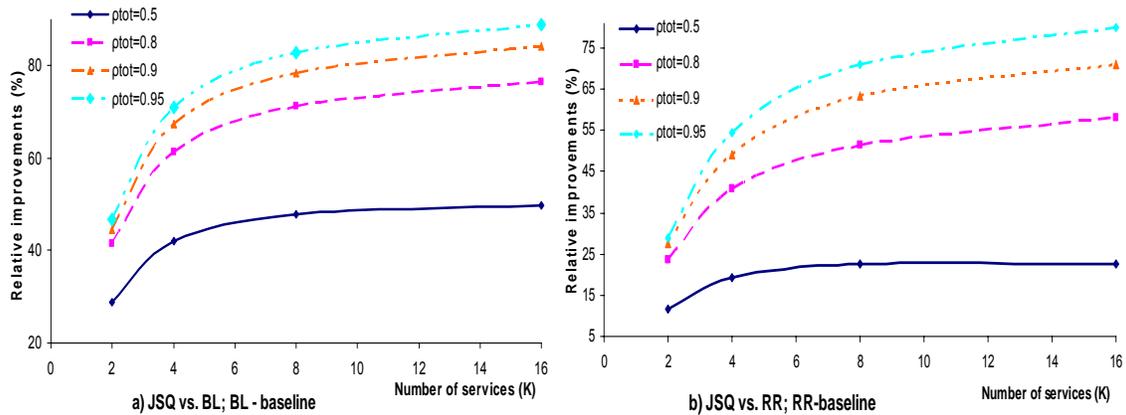


Figure 4. Relative comparison of mean response times between JSQ and BL (left) and JSQ and RR (right) strategies for different number of services K and different values of ρ_{tot} . The system state information is up-to-date and only foreground traffic is present in the system.

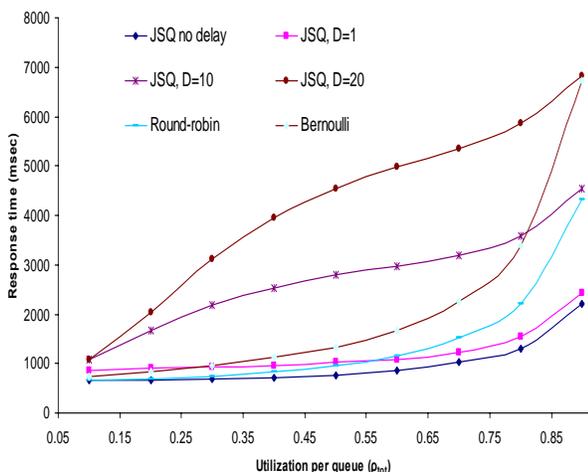


Figure 5. Comparison of mean response times for the following dispatching strategies: BL, RR, JSQ with up-to-date system state information, and JSQ when system state information (queue lengths) is obtained with period $\Delta = D \cdot \frac{1}{\lambda_{tot}}$, where $D \in \{1, 10, 20\}$. Background traffic is not present and the number of services $K = 4$.

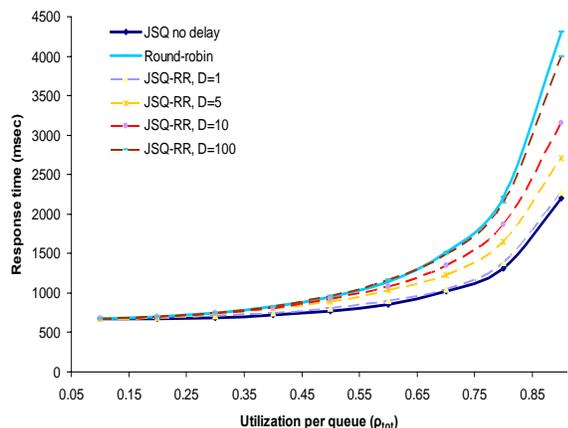


Figure 6. Comparison of mean response times for the following dispatching strategies: RR, JSQ with up-to-date system state information and JSQ-RR when system state information (queue lengths) is obtained with period $\Delta = D \cdot \frac{1}{\lambda_{tot}}$, where $D \in \{1, 5, 10, 100\}$. Background traffic is not present and the number of services, $K = 4$.

Our simulations and analysis are directed to answering the question of the impact of the background traffic to the response times. The simulations results are shown in Figure 7 for $K = 4$ services and BL, RR, and JSQ strategies. Since the system state information is assumed to be instantaneously available, JSQ-RR is identical to JSQ, and therefore not shown. We have recorded the response times of the foreground requests only. For given utilization per queue ρ_{tot} , and dispatching strategy, foreground traffic percentage of ρ_{tot} has been varied from as little as 10%

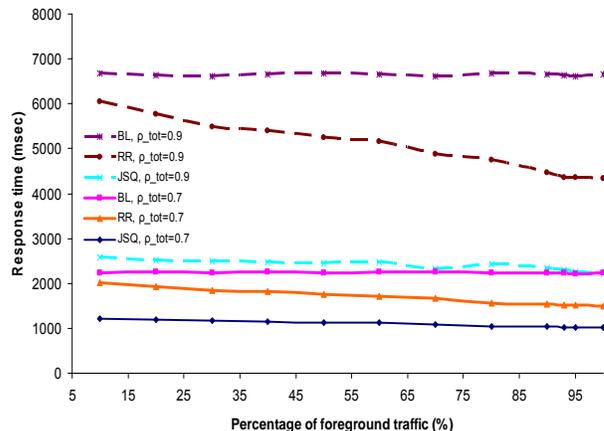


Figure 7. Response times for BL, RR and JSQ strategies for scenario with background traffic and no information delay. Utilization per queue ρ_{tot} is 0.7 or 0.9, and number of services $K = 4$.

(i.e., 90% background traffic) to 99% (i.e., 1% background traffic). Apart from the case of the Bernoulli dispatching strategy when response times are constant, as expected, from Figure 7 it follows:

- In case of the Round-Robin strategy response times decrease as the percentage of foreground traffic increases. It seems that response times dependency from the given percentage is linear.
- In case of the JSQ strategy response times show linear non-increasing dependency from the given percentage of foreground traffic. The decrease of the response time is limited by 15% for the considered cases. It seems that the JSQ strategy is not much sensitive to the background traffic.

The intuitive explanation for the decreasing nature of response times in case of RR and JSQ strategies may be given as the following – the response time in the case of these two strategies is biggest for the smallest percentage of foreground traffic, due to the fact that only foreground traffic is “intelligently” assigned to one of the queues.

Response time for JSQ with low foreground traffic load: Let us now consider the situation where the foreground traffic constitutes only a small percentage of the total traffic. We will analyse the mean response time of a tagged foreground traffic arrival. According to the JSQ policy this arrival will be dispatched to the queue (out of K queues) with the smallest length. Since the foreground traffic is negligible and the background traffic arrival processes are i.i.d., the random processes representing the queue lengths are also independent from each other and behave as the queue length of an M/M/1 PS queueing model with load ρ_{tot} . The queue length distribution for this model is geometric with parameter ρ_{tot} , see [22]. Hence, the probability that the queue selected for the tagged foreground job contains n

(background) jobs is given by:

$$\Pr\{n \text{ jobs in selected queue}\} = (1 - \rho_{\text{tot}}^K) (\rho_{\text{tot}}^K)^n.$$

Once the tagged arrival is placed to a particular queue, that queue further behaves as an “ordinary” M/M/1 PS queue with utilization $\rho_{\text{tot}} = \frac{\lambda_{\text{BT}}}{\mu}$, as the foreground traffic is negligibly small.

Now, let us denote by $X_n(\tau)$ the random variable whose distribution is that of the “delay” experienced by the tagged arrival if it would have service requirement τ and arrives when there are n background jobs in the queue. The total time spent in the system for the tagged arrival (i.e., response time) is then $X_n(\tau) + \tau$.

From the detailed analysis of the M/M/1 PS queue in [7], it follows that (cf. Eq. (33) in [7]):

$$E\{X_n(\tau)\} = \frac{\rho_{\text{tot}}\tau}{1 - \rho_{\text{tot}}} + [n(1 - \rho_{\text{tot}}) - \rho_{\text{tot}}] \cdot \frac{1 - e^{-(1-\rho_{\text{tot}})\mu\tau}}{\mu(1 - \rho_{\text{tot}})^2}.$$

Since the r.v. $X_n(\tau)$ is conditioned by n , the mean $E\{X(\tau)\}$ is given by the following equation

$$E\{X(\tau)\} = \sum_{n=0}^{\infty} (1 - \rho_{\text{tot}}^K) (\rho_{\text{tot}}^K)^n \cdot E\{X_n(\tau)\},$$

which leads to

$$E\{X(\tau)\} = \frac{\rho_{\text{tot}}\tau}{1 - \rho_{\text{tot}}} + \frac{\rho_{\text{tot}}^K - \rho_{\text{tot}}}{1 - \rho_{\text{tot}}^K} \cdot \frac{1 - e^{-(1-\rho_{\text{tot}})\mu\tau}}{\mu(1 - \rho_{\text{tot}})^2}.$$

The overall mean response time for the tagged arrival is given by

$$\text{RT} = \frac{1}{\mu} + \int_0^{\infty} E\{X(\tau)\} d(1 - e^{-\mu\tau}),$$

which finally gives

$$\text{RT} = \frac{1}{\mu} + \frac{\rho_{\text{tot}}}{\mu(1 - \rho_{\text{tot}})} + \frac{\rho_{\text{tot}}^K - \rho_{\text{tot}}}{1 - \rho_{\text{tot}}^K} \cdot \frac{1}{\mu(1 - \rho_{\text{tot}})} \cdot \frac{1}{2 - \rho_{\text{tot}}}.$$

This equation gives a surprisingly simple relationship between the response time for the foreground traffic, the number of services K , the utilization per queue ρ_{tot} and the mean of the foreground job sizes $\frac{1}{\mu}$.

These formulae have been deduced under the assumption that foreground traffic intensity is negligible compared to background traffic. Inspired by the numerical results in Figure 7 we investigated whether this response time formula could be used as an approximation for larger values of the percentage of the foreground traffic. A first comparison between our approximate formula and simulations, taking the simulations as the baseline, is given in Table I. The comparison indicates that:

- As expected, for a fixed number K of concrete services, the difference between our analytical results and simulation increases when the percentage of foreground traffic becomes larger. This is because our formula has

been deduced under the assumption that there is only one foreground traffic arrival.

- Roughly speaking, the error of our approximate formula increases as the number of services K increases (and all other parameters remain the same).
- The relative difference between our formula and simulation increases when ρ_{tot} increases and all other parameters remain the same

TABLE I
RELATIVE COMPARISON BETWEEN THE RESPONSE TIMES OBTAINED BY SIMULATIONS AND RESPONSE TIMES CALCULATED USING THE FORMULA.

FG traffic (%) →	$K = 2$		$K = 4$		$K = 8$	
	5	10	5	10	5	10
$\rho_{\text{tot}} = 0.5$	0.1%	0.19%	0.5%	1.6%	1.8%	1.5%
$\rho_{\text{tot}} = 0.7$	1.1%	1.7%	1%	1.3%	2.4%	8.2%
$\rho_{\text{tot}} = 0.9$	1.7%	4.6%	5.2%	8.6%	5.0%	13.5%

D. Scenarios with background traffic and stale system state information

For these scenarios we have conducted simulations in order to investigate which factor has more impact to the response time: delayed system state information or background traffic.

The simulation results presented in Figure 8 for the JSQ-RR strategy, apply to the case when ρ_{tot} is fixed at 0.7 and the number of services $K = 4$. Results are shown for four different values of the parameter D representing the system state information delay: $D \in \{1, 2, 5, 10\}$. As for the case with up-to-date system state information ($D = 0$) considered in the previous subsection, we see that the response time as function of the percentage of foreground traffic has a decreasing trend. Obviously, when background traffic diminishes, the response time approaches the values obtained for the scenarios without background traffic considered in subsection IV-B. However, all together, it is hard to determine from this figure which of the two factors has predominant influence on the response time.

In order to investigate whether delayed system state information or intensity of the background traffic has more impact to the system performance, we compare results from Figure 8 ($RT_{\text{BG}+\Delta}$) to results when only stale information is present (RT_{Δ}). The comparison is presented at Figure 9 and represents the ratio $r = \frac{RT_{\text{BG}+\Delta}}{RT_{\Delta}}$ for different values of the system state information delay parameter D . The ratio r is lower bounded by 1, and when $r \rightarrow 1$ delay has more influence on $RT_{\text{BG}+\Delta}$ than background traffic. The following conclusions can be made from Figure 9:

- The larger D , the more influence has the background traffic on $RT_{\text{BG}+\Delta}$. Suppose that the percentage of the foreground (background) traffic in the system is fixed. As D increases, the interval when state information is collected becomes larger. The larger the interval, the

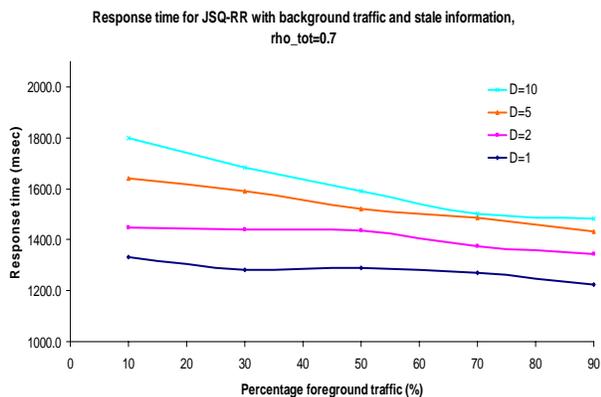


Figure 8. Response times for JSQ-RR strategy in case of the scenario with background traffic and stale information, with parameter $D \in \{1, 2, 5, 10\}$. Utilization per queue, ρ_{tot} is 0.7, the number of services is $K = 4$.

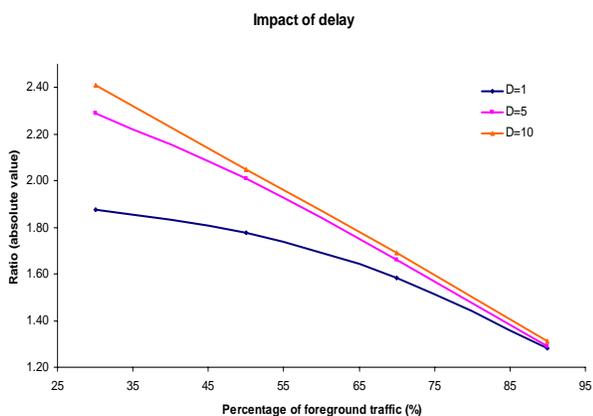


Figure 9. Ratio of response times between the system with background traffic and stale information and the system with stale information only. The dispatching strategy is JSQ-RR. Utilization per queue is $\rho_{tot} = 0.7$ and the number of services is $K = 4$.

more background traffic arrivals to a queue between two state information updates. The response time of the tagged foreground arrival will therefore be influenced by more background arrivals.

- For smaller values of parameter D , the relative change of ratio r is smaller. For example, when $D = 1$ the ratio changes from 1.88 (30% foreground traffic) to 1.28 (90% foreground traffic), compared to change from 2.41 to 1.31 when $D = 10$, respectively. This means that absolute influence of background traffic is smaller for smaller values of D . The smaller the period of the system state information update, less background traffic arrivals are probable within one such period.

V. CONCLUSION AND FUTURE WORK

In this paper, we have investigated the performance potential of dynamic, run-time web service selection within SOA, under various assumptions regarding the available system state information and/or presence of background traffic. Using simulation and analytical modelling it has been shown that, compared to static (a-priori) service selection, considerable performance improvements are possible, even when the state information is stale and/or background traffic is present. These improvements result from exploiting workload fluctuations that occur at relatively small time scales mainly caused by the random behaviour of potential clients.

The main results of the paper could be summarized as the following:

- 1) Quantification of the achievable gain versus the number K of pre-selected concrete services by a fair comparison with respect to the base case of static web service selection. For relatively small numbers of K , e.g., $K = 4$ or $K = 8$, significant response time reductions are obtainable.
- 2) Quantification of the achievable gain in terms of response times when background traffic is present at the pre-selected concrete services for different dispatching strategies. We show that the response time performance of JSQ is quite robust with respect to the presence of background traffic. An insightful approximate formula for the response time under the JSQ dispatching strategy is derived for cases where the background traffic is dominant.
- 3) Quantification of the achievable gain in terms of response times in case of delayed state information. A stateless dispatching algorithm such as RR always improves upon the base case. Statefull dispatching algorithms such as JSQ should be carefully applied as can potentially perform worse than the base case when delay is present. However, a combination of RR and JSQ, referred to as JSQ-RR, always improves on RR and hence the base case, even if the delay of the system status updates tends to infinity. In fact, the response time performance of JSQ-RR is upper bounded by the performance of RR, and lower bounded by JSQ.

Nevertheless, the promising results raise several research questions still to be answered, e.g.:

- What is the performance under more general assumptions regarding the requests' arrival processes and their service requirements?
- What is, eventually, the impact of the resulting overhead (due to making the required system state information available) on the performance?
- What is the performance of alternative runtime dispatching strategies that don't introduce additional overhead, e.g. dispatching strategies based on response times from previously assigned jobs instead of explicit

(“stale”) system information?

- What is the performance of the observed dispatching strategies when the service composition comprises multiple abstract services?

ACKNOWLEDGMENT

This work has been carried out in the context of the IOP GenCom project Service Optimization and Quality (SeQual), which is supported by the Dutch Ministry of Economic Affairs via its agency Agentschap NL.

REFERENCES

- [1] E. Altman, U. Ayesta, and B.J. Prabhu. Optimal load balancing in processor sharing systems. In *Value Tools '08*, pages 1–10, ICST, 2008.
- [2] F. Bonomi. On job assignment for a parallel system of processor sharing queues. *IEEE Trans. Comput.*, 39(7):858–869, 1990.
- [3] S. Borst. Optimal probabilistic allocation of customer types to servers. *Proc. ACM SIGMETRICS*, pages 116–125, ACM, 1995.
- [4] G. Canfora, M. Di Penta, R. Esposito, and M.L. Villani. An approach for QoS-aware service composition based on genetic algorithms. *Gecco Proceedings*, 2005.
- [5] J. Cao and C. Nyberg. An approximate analysis of load balancing using stale state information for servers in parallel. *Proc. Second IASTED International Conference on Communications, Internet, and Information Technology*, pages 17–19, 2003.
- [6] Y.-C. Chow and W. H. Kohler. Models for dynamic load balancing in a heterogeneous multiple processor system. *IEEE Trans. Comp.*, 28(5):354–361, 1979.
- [7] D. G. Coffman, N. R. Muntz, and H. Trotter. Waiting time distributions for processor-sharing systems. *Journal of the ACM*, 17(1):123–130, 1970.
- [8] Université de Montréal, Département d’Informatique et de Recherche Opérationnelle (DIRO). Stochastic simulation in java, SSJ. <http://www.iro.umontreal.ca/~simardr/ssj/indexe.html>, last accessed January 2011.
- [9] L. Flatto and H. P. McKean. Two queues in parallel. *Comm. Pure and Applied Mathematics*, 30:255–263, 1977.
- [10] V. Gupta, M. Harchol-Balter, K. Sigman, and W. Whitt. Analysis of join-the-shortest-queue routing for web server farms. *Performance Evaluation*, 64(9–12):1062–1081, 2007.
- [11] G.J. Hoekstra, Rob van der Mei, Y. Nazarathy, and A.P. Zwart. On the sojourn time tails of a file-splitting processor sharing network. *Proc. NET-COOP, LNCS Vol. 5894*, 2009.
- [12] S. Hwang, E. Lim, C. Lee, and C. Chen. Dynamic web service selection for reliable web service composition. *IEEE Trans. Services Comp.*, 1(2):104–116, 2008.
- [13] G. A. Lewis, E. Morris, D. B. Smith, S. Simanta, and L. Wrage. Common misconceptions about service-oriented architecture. *Proc. Sixth International IEEE Conference COTS-Based Software Systems*, pages 123–130, IEEE, 2007.
- [14] Y. Liu, A.H.H. Ngu, and L. Zeng. QoS computation and policing in dynamic web service selection. *Proc. 13th International World Wide Web Conference*, pages 66–73, 2004.
- [15] Z. Liu, N. Niclausse, C. J. Vilanueva, and S. Berbier. Traffic model and performance evaluation of web servers. *Performance Evaluation*, 46(2–3):77–100, 2001.
- [16] M. Mitzenmacher. How useful is old information? *IEEE Trans. Parallel and Dist. Syst.*, 11(1):6–20, 2000.
- [17] R. D. Nelson and T. K. Philips. An approximation to the response time for shortest queue routing. *ACM Performance Eval. Review*, 17(1):181–189, 1989.
- [18] R. D. Nelson and T. K. Philips. An approximation to the response time for shortest queue routing with general interarrival and service times. IBM T.J. Watson Research Lab Technical Report RC15429, 1990.
- [19] L. D. Ngan, M. Kirchberg, and R. Kanagasabai. Review of Semantic Web Service Discovery Methods, *Proc. IEEE 6th World Congress on Services*, pages 176–177, IEEE, 2010.
- [20] A. Padovitz, S. Krishnaswamy, and S. Loke. Towards efficient selection of web services. *Second Int. Joint Conf. on Autonomous Agents and Multi-Agent Systems*, July 2003.
- [21] S.R. Ponnekanti and A. Fox. Sword: A developer toolkit for web service composition. *Proc. 11th International World Wide Web Conference*, pages – 2002.
- [22] Henk Tijms. *A First Course in Stochastic Models*. John Wiley and Sons, Chichester, England, 2003.
- [23] T. Yu, Y. Zhang, and K.-J. Lin. Efficient algorithms for web services selection with end-to-end QoS constraints. *ACM Trans. Web*, 1(1):p.6-es, 2007.
- [24] L. Zeng, B. Benatallah, A.H.H. Ngu, M. Dumas, J. Kalaganam, and H. Chang. QoS-aware middleware for web services composition. *IEEE Trans. Softw. Eng.*, 30(5):311–327, 2004.

The Strategic Role of IT: An Empirical Study of its Impact on IT Performance in Manufacturing SMEs

Louis Raymond
Université du Québec
Trois-Rivières, Canada
louis.raymond@uqtr.ca

Anne-Marie Croteau
Concordia University
Montréal, Canada
anne-marie.croteau@concordia.ca

François Bergeron
Télé-Université
Québec, Canada
bergeron.francois@teluq.uqam.ca

Abstract — The business value of Internet and Web applications for small- and medium-sized enterprises (SMEs) is dependent upon how such applications “fit” with the strategic orientation of these firms. Based on the strategic alignment of information technology (IT), this study uses a systemic approach to test the contribution of three predictors of IT performance in an organization: the strategic role of IT as well as the sophistication of the management and the use of IT. A multivariate mediation perspective is used to conceptualize alignment. The results of an empirical investigation of 44 manufacturing SMEs establish an important mediator effect of IT management and IT usage sophistication between the strategic role of IT and IT performance.

Keywords - *IT sophistication; IT performance; SME; strategic alignment; strategic role; e-business applications*

I. INTRODUCTION

The current economic context is marked by a considerable expansion of electronic markets. The need for information technology (IT) comes with the ever increasing demand for digital information. The impact on business is tremendous, especially for small- and medium-sized enterprises (SMEs) who need to invest in systems with the ability to store, process, and generate data stemming from its dealings with various business partners. For many small companies, the need for IT is necessary to ensure their survival and competitiveness [25], and to enable their innovation capabilities [15]. While it is evident, IT systems have substantially improved the manufacturing process and productivity, they have also allowed for more organizational flexibility by transforming workers time and space, and reinventing the internal and external mode of business organization [14].

With the pervasive arrival of IT and the opening of the global marketplace, SMEs have been subject to rapid change and extreme instability. The SME is obliged to integrate IT to ensure its competitiveness and survival. IT can play an important role in a company's performance and its ability to respond effectively to the changing needs of the market, therefore special attention and research is needed [33].

And given the inherent fragile nature of the SME, IT plays an increasingly strategic role in creating new challenges [11]. New online competitors easily enter new or existing markets, customers are more informed and more demanding because they can compare features and prices of products through the Web; also, the changing needs and wants of the market often render recent IT investments obsolete [36]. As a result of substantial investments by many SMEs into IT, it is essential to foresee the threats and opportunities that are inherent in the technology, to discover the mechanisms that manage and drive these technologies, and to analyze the impact in terms of cost-effectiveness and profitability for these enterprises [22].

The increased strategic nature of the role of IT in the organization may give rise to management problems and IT use, not only at a technical level, but also at strategic and organizational levels [5]. It is therefore important for SMEs to better understand how their investments in IT, coupled with an increased understanding of their management practices and how they use the technology, provides the most benefits [16]. This research thus aims at better understanding the impact of the strategic role of IT, IT management sophistication, and IT usage sophistication upon IT performance in manufacturing SMEs, by answering the following research question: *To what extent and in what manner do the strategic role of IT, the sophistication of IT management, and the sophistication of IT usage contribute to the performance of IT in manufacturing SMEs?*

We first present the theoretical background of the research, followed by the research model, and the method by which 44 French manufacturing SMEs were empirically studied in order to answer the research question. Next, the results are presented and discussed. We further identify the study's implications and limitations, and conclude with future research.

II. THEORETICAL BACKGROUND

The study's theoretical background is founded on the concept of *strategic alignment* at the core of the

strategic paradigm in information systems research. It was first defined in terms of its impact on organizational performance rather than on the performance of IT. However, there is no doubt the concept is still one of the most important fundamental bases of our understanding of the strategic role of IT and IT performance in organizations. According to Henderson and Venkatraman [18], strategic alignment is based on the assumption of a dynamic and coherent IT (strategy and infrastructure), the company's business strategy and process development, would have an impact on performance and thus its competitiveness. An enterprise should synchronize its business and its technology sectors, as well as at the strategic and operations levels. As presented in Fig. 1, Henderson and Venkatraman's [18] model is based on a systemic approach, emphasizing the importance of aligning internal and external business activities in order to improve organizational performance and predetermined strategic objectives.

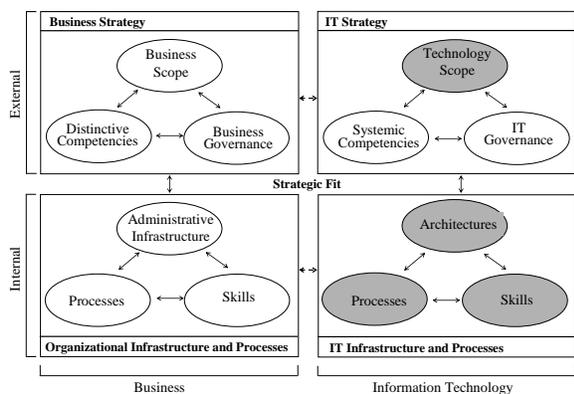


Figure 1. Adaptation of Henderson and Venkatraman's [18] IT alignment framework

This model of IT strategic alignment has been tested in various ways over the past two decades, often by exploring in more detail some aspects of each dimension described above. However, few studies have taken into account other factors such as organizational context (business strategy, organizational structure), the environmental context (industry, firm size), and the technological context (technology solutions, IT management) [39]. It has been highlighted that a close link between business strategy and IT strategy contributes to both IT performance and organizational performance [8, 12]. Despite previous empirical studies that have allowed us to better understand the contexts in which the alignment of IT contributes or not to organizational performance, many aspects remain unexplored, including alignment at the technological level [2].

This study proposes a research model for ascertaining in what manner IT “works” in SMEs, specifically the strategic role of IT, and the sophistication of IT management and IT usage. IT performance is seen here as a result of direct or “proximal” strategic alignment of IT [8], while organizational performance would be considered rather as an indirect consequence or “distal” of this alignment. Returning to Fig. 1, the shaded sections of Henderson and Venkatraman's [18] model are the basis of the research model used for the current study.

A. Strategic Role of IT

According to Powell and Dent-Micallef [32], IT human resources and the organizational structure of the business must complement one other in a way that creates intrinsic benefits that results in a significant and distinctive performance compared to other sectors in the company. Thus, more and more emphasis is needed to optimize the use and management of IT based on the internal characteristics of the company as well as its strategic profile, size, in-house expertise, as well as its managerial, technological, and functional capacity.

Certain researchers have explored the idea of the evolution of IT usage, namely Ward, Taylor and Bond [43] who observed that the strategic role of IT is developed over three major periods in order to support the business while throughout its growth: (1) The period of developing data processing standards and automating repetitive tasks, thereby improving operational efficiency, (2) The period of managing information systems, designed to improve management efficiency by producing concrete information that will be used to better manage and control the firm, (3) The period of strategic information systems enabling the company to better position itself in its market segment.

As per the model proposed by Philip and Booth [31], each organization has its own expectations with regards to IT that is dependent upon its skills and capabilities to align the technology with its strategic objectives. According to this model, information systems can play five potential roles in the enterprise: (1) Survival is the most important role played by IT in an organization. The goal of IT is to achieve greater control over management and can be used to understand day-to-day administrative and production tasks in order to achieve operational performance and cost reduction. (2) Resources: this model compares the company to a waterway, receiving a stream of resources such as materials and services of other companies, and issuing products and services that can be used by other third parties. (3) Competitive advantage: IT moves from the simple role as facilitator of obtaining resources to one of fully exploiting potential of resources to gain a competitive advantage. Creativity and innovation are

essential ingredients to this approach. (4) Value analysis service: rethinking business processes by reengineering them to improve the company's competitiveness and flexibility, taking into account the rapid changes in the environment. (5) Cyberspace: It is in cyberspace that virtual organizations build relationships with suppliers, consumers, and other organizations. This type of structure is very flexible, innovative, and provides very personalized service.

B. IT Sophistication

The evolution of the strategic role of IT is closely linked to IT sophistication because it reflects the way IT is managed and used by the company. IT sophistication is explained by the way IT falls into line with the firm's strategic objectives [34]. The concept of IT sophistication and its measurement were first defined and validated by Raymond, Paré and Bergeron [35], to be subsequently used by other researchers [10, 19, 29]. IT sophistication refers to the nature, complexity, and interdependence of IT use and how it is managed within the organization. IT sophistication management includes managerial and functional sophistication on the one hand, while the use of IT sophistication includes informational and technological sophistication on the other hand.

Managerial sophistication takes into account the mechanisms used to plan, monitor, and assess current and future applications. Within the context of the SME, the sophistication of planning is demonstrated by the degree of formalism of the company's processes and the level of alignment with the organizational goals. The availability of written documents, standards, and measures for purposes of security and confidentiality clearly reveal a managerial sophistication. This dimension may also contain aspects related to the accomplishment of business objectives relating to the adoption of IT, the degree of formalization of the adoption process as well as managerial involvement in this process. Also related to managerial sophistication are the presence of external consultants, the initial investment, and the annual budget allocated to develop and operate IT applications.

Functional sophistication refers to the location and functional autonomy of IT within the organization. The number of internal specialists within IT function is an indicator insofar as it refers to the number and nature of the tasks to be completed by the IT function or by the amount of persons responsible for IT within the organization. Based on the hierarchical level of the organization (operational, administrative, strategic), sophistication can be characterized by the proportion of IT applications at each of these decision-making levels.

Informational sophistication refers to the nature, both transactional and managerial, of the applications

portfolio [4]. Another aspect of informational sophistication is the degree in which applications are integrated in the SME; this element can be characterized by the presence of a central database shared by the implantation of a LAN, by a fluid circulation of information and sharing of resources, and by the implementation of an integrated management software package (ERP). The importance of the ERP system at this level would allow businesses to share information with its partners in real-time [7].

Technological sophistication reflects the number or variety of IT used by the SME in several areas such as CAD/CAM, internal networking, and external networking [23]. In each of these areas, the number of hardware and software integrated into the system is counted based on the degree of complexity and type of technology used by the organization. The technological dimension considers the nature of the hardware and development tools used by the SME. The number of workstations is a first indicator of IT sophistication in addition to decentralized hardware within the organization. A second indicator is the diversity of programming languages and development tools used. The sophistication of the man-machine interface can also be considered as a criterion. Finally, dominant treatment modes as well as the types of operation preferred by the SME are added to the mass of indicators of technological sophistication. The adoption of an ERP system can be considered as a perfect example of technological sophistication within the SME, who are attempting to prevent the consequences of technological obsolescence by more sophisticated integrated applications such as FMS [40].

C. IT Performance

Assessing the performance of the IT function in an organization is not a simple task [26]. In a process evaluation of IT costs, Keen [21] proposed taking into account various elements such as the technical obsolescence of software, the declining cost of work units and operating software, development flows, and operating costs. IT takes the form of other assets and is also subject to devaluation or replacement. These items could be used to quantitatively assess the standard IT budget including machines, applications, and services. However none of them consider organizational transition costs related to learning, reticence, stress, fault, change internal reporting, information loss, and additional migration costs.

Benefits gained from IT remain very complex to identify specifically in relation to profitability studies. In addition, quantifying benefits from organizational change, improved customer follow-up, or even an improvement of internal and external communication, are a challenge for a number of enterprises. In fact,

there is the dilemma of quantifying qualitative and intangible inputs with indicators for legacy assets. Some companies choose to go beyond this notion of IT profitability, by categorizing them as machinery and equipment for production, such as furniture, office or as general operating costs resulting in a shortsighted view on IT investment. Others use indicators to measure operational performance (quality indicators, satisfaction surveys), technical (referencing, application availability, application evolution), or users (number of users of a system, effective consultations frequency). DeLone and McLean [13] developed a model that IT success can be measured via six dimensions; quality of the system, quality of the information, usage, user satisfaction, individual impact, and organizational impact. User satisfaction remains however one of the most important measures of success and most recognized in IT [37]. It has been demonstrated that the quality of the system, the quality of the information and the usefulness of applications point to, in large part, the satisfaction of users.

III. RESEARCH MODEL

As presented in Fig. 2, the research model is based upon a conceptualization of the strategic alignment of IT proposed by Henderson and Venkatraman [18], more specifically the alignment between the IT strategy and the IT infrastructure and processes that is deemed to have a positive impact upon the performance of IT in manufacturing SMEs. The IT strategy is as the strategic role attributed to IT by the SME’s leader, whereas the IT infrastructure and processes are as the firm’s sophistication in both managing and using IT. Testing this model should help us answer the following research question: *To what extent and in what manner do the strategic role of IT, the sophistication of IT management, and the sophistication of IT usage contribute to the performance of IT in manufacturing SMEs?*

As shown in the research model, the strategic role of IT is an independent construct directly related to the dependent construct, i.e. IT performance. The impact of the strategic role of IT will also be felt by the sophistication of IT management and IT usage. This research model aims to explain IT performance in a novel way by focusing on the strategic role of IT while taking into account the sophistication level of IT deployed in manufacturing SMEs. It is for this reason that the IT sophistication concept [35] is mobilized here, that is, IT management sophistication on one hand, and IT usage sophistication on the other hand. The first hypothesis is in line with the main proposition found in previous conceptualizations of the strategic alignment of IT on the basis of the evolution of information technology’s role in organizations [31, 42].

Its distinction and contribution however lie in the choice of IT performance (or organizational IS effectiveness) rather than organizational performance as the outcome of such alignment.

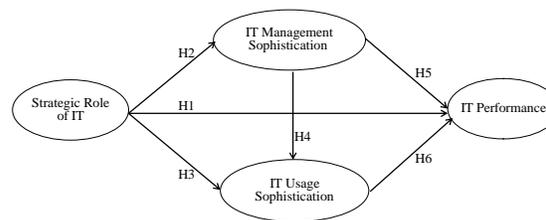


Figure 2. Research model

It had been previously noted that the strategic role played by IT in organizations could only be ascertained if one took into account their IT management and usage characteristics. Now the notion of IT sophistication effectively reflects how IT are managed and used within organizations [43]. Hence the second hypothesis assumes the more strategic the role played by information technology in the organization, the greater the presence of its IT function. Following Henderson and Venkatraman [18], it is presumed that in manufacturing SMEs, the strategic importance of IT will be reflected in the IT resources and capabilities developed by the IT function. The third hypothesis reflects the premise that users will be more satisfied with the applications implemented and with the quality of information output if the SME’s leadership views IT as a strategic necessity or as a source of competitive advantage. Here, the notion of “top-management support” as a determinant of IS success would take on added importance in small business [37].

The fourth hypothesis assumes a certain hierarchy in the evolution of IT, as previously indicated, i.e. this technology must be effectively managed and deployed in the SME if it is to be appropriated and effectively used by employees. This is basically in line with DeLone and McLean’s [13] updated IS success model in which system usage and user satisfaction are dependent upon the quality of the system, the information output, and the service provided by the IT function [30]. The fifth hypothesis proposes that the performance of IT improves when the sophistication of the management of IT increases [27]. As noted by Philip and Booth [31], “sustainable advantage depends on the ability to manage the IS resources effectively on an ongoing basis”. The last hypothesis similarly

proposes that IT performance improves with more sophisticated usage of IT [35], this being in line again with DeLone and Mclean’s [13] IS success model.

In summary, the following hypotheses are tested:

- H1: The more strategic the role played by IT, the higher the performance of IT.
- H2: The more strategic the role played by IT, the greater the IT management sophistication.
- H3: The more strategic the role played by IT, the greater the IT usage sophistication.
- H4: The greater the IT management sophistication, the greater the IT usage sophistication.
- H5: The greater the IT management sophistication, the higher the performance of IT.
- H6: The greater the IT usage sophistication, the higher the performance of IT.

IV. METHOD

Secondary data was provided by a database created by a university research center for benchmarking purposes and containing information of 44 French manufacturing SMEs. For the study’s purpose, a SME is defined as having between 10 and 299 employees, the median size of the sampled firms being 38 employees. The industrial sectors represented include metals (27%), food and beverage (16%), wood (9%), plastics (9%), textile (7%), minerals (5%), electronics (2%) and others (25%).

A. Data Collection

This database was created in collaboration with business owners that belong to chambers of commerce in Midi-Pyrénées region, by asking the management team and IT manager to answer a questionnaire on the firm’s strategic orientation, practices, and performance with regard to information technology and e-business, broken down by the main business functions of the SME, namely operations and production, sales and marketing, and accounting, finance and HRM. In exchange for this information, the firm was provided with an overall diagnostic of its situation relative to the management and performance of its information technology.

B. Measures

In view of Henderson and Venkatraman’s [18] framework on which this research is based, fit or alignment between the strategic role of IT and the sophistication of IT management and usage in the firm is ascertained here from a “fit as mediation” perspective [41]. First, the extent to which IT plays a strategic role in the SME was measured through a self-typing approach based on Venkatraman’s [42] and Philip and Booth’s [31] stage models, by asking the chief executive to answer the following question (statements

were coded from 1 to 4 in order of increasing strategic importance):

Indicate among the following statements the one that best defines your understanding of the strategic role that is assigned to information technology-based applications (ITApps) in your firm (choose one statement)?

1. ITApps should allow us to improve our managerial control and our production monitoring.	<input type="checkbox"/>
2. ITApps should insure greater operational flexibility and better response to our customers’ needs.	<input type="checkbox"/>
3. ITApps should facilitate and accelerate the development of new products, and allow us to increase our market share.	<input type="checkbox"/>
4. ITApps should allow us to integrate our business and production processes, and to improve exchanges with our business partners.	<input type="checkbox"/>

The measures of IT management sophistication, in terms of managerial and functional sophistication, and of IT usage sophistication, in terms of informational and technological sophistication, emanate from constructs developed, validated, and used in previous research [29, 35]. IT performance is measured by the level of attainment of the benefits associated with four types of IT-based applications (accounting-finance-HRM, logistics-production-distribution, marketing-sales-customer service, e-business-Internet-Web), thus following a process-based approach wherein the respondents evaluate the “business value” of IT for their firm [38, 40]. A list of expected benefits specific to each type of application (e.g. “increase flexibility”, “improve customer service”, “facilitate the recruitment of personnel”) is presented to the manager (CEO or CFO, operations manager, sales and marketing manager, and IT manager) who must indicate on a 5-point scale the extent to which the applications implemented contribute to the attainment of these benefits.

V. RESULTS

Structural equation modeling was used to validate the research model. To this effect, the PLS technique was chosen for its robustness, more precisely its capacity to handle small samples and formative measurement models in comparison to covariance structure analysis techniques such as Lisrel, EQS and Amos [17].

A. Measurement Model

Given their composite and multidimensional nature, the research constructs are modeled as being “formative” rather than “reflective” [9]. Such a construct is composed of many indicators that each

captures a different aspect; hence changes in these indicators bring or “cause” change in their underlying construct [24]. IT management sophistication is thus modeled as a second-order formative construct from two sub-constructs, namely managerial sophistication and functional sophistication. As presented in Table 1, each of these sub-constructs is in turn composed of six and two formative measures respectively, a functional sophistication and managerial sophistication score being obtained from the factor scores determined by a principal components analysis. Given that this analysis produced two components for managerial sophistication, a single score was obtained by averaging the two factor scores.

The reliability of a formative construct, as opposed to a reflective one, is confirmed by the absence of multicollinearity between its measures or indicators [28]. Formative indicator validity is confirmed by a weight that is significant and not less than 0.1 [20], as confirmed in Fig. 3. Discriminant validity of a formative construct is confirmed by it sharing less than 50% variance with any other construct, whereas nomological validity is confirmed when the construct’s hypothesized links with other constructs are significantly greater than zero and in the expected direction [1].

TABLE 1. PRINCIPAL COMPONENTS ANALYSIS OF IT MANAGEMENT SOPHISTICATION

indicator	factor	Funct. Soph.	Man. Soph. ^a	Man. Soph. ^b
Functional Sophistication	designated manager for IT	.91	-	-
	org. level of the IT function	.91	-	-
Managerial Sophistication	IT development	-	.79	-
	IT evaluation	-	.68	-
	user participation	-	.75	-
	external consultants	-	.58	-
	IT resources&competencies	-	-	.93
	IT support & appropriation	-	-	.95

^aIT management practices

^bIT management capabilities

In similar fashion, IT usage sophistication is modeled and measured from two sub-constructs, namely informational sophistication and technological sophistication. As presented in Table 2, each sub-construct is in turn composed of six and three indicators respectively. The reliability and validity of the IT usage sophistication construct was similarly confirmed. As to the IT performance construct, it is composed of four measures, that is, the average benefits obtained from each type of IT-based application. One may note again that there is no multicollinearity among these last

formative measures, the highest correlation among them being equal to 0.19 (p > 0.1), with all four regression weights being greater than 0.1 (Fig. 3), thus showing adequate reliability and validity.

TABLE 2. PRINCIPAL COMPONENTS ANALYSIS OF IT USAGE SOPHISTICATION

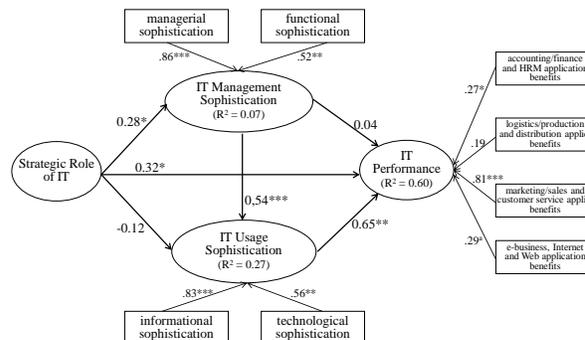
indicator	factor	Tech. Soph.	Inf. Soph. ^a	Inf. Soph. ^b
Technological Sophistication	uses of IT	.90	-	-
	uses of e-bus/Internet/Web	.81	-	-
	quality of IT security	.50	-	-
Informational Sophistication	accounting/fin./HRM apps	-	.78	-
	logistics/prod./distrib. apps	-	.60	-
	mark./sales/cust. serv. apps	-	.74	-
	ERP system modules	-	.69	-
	information output quality	-	-	.93
	user-system interaction qual.	-	-	.95

^aextensiveness of IT usage

^bquality of IT usage

B. Test of the Research Model

The research hypotheses were tested by evaluation the direction, value, and level of significance of the path coefficients estimated by PLS, as presented in Fig. 3.



Nota. Significance levels were obtained by bootstrapping. *p < 0.1 **p < 0.05 ***p < 0.001

Figure 3. Results of testing the research model

A positive and significant path coefficient ($\gamma_1 = 0.32$; $p < 0.05$) confirms the first research hypothesis, that is, the more strategic the role played by IT in the manufacturing SME, the greater its IT performance. Moreover, if one removes the effect of IT management and IT usage sophistication upon IT performance, the strategic role of IT still explains 25% of the variance in this same performance. The benefits obtained from marketing, customer service, and e-business applications thus flow directly from a vision of IT as a

mean for the SME to develop its products and markets, to integrate its production processes, and to improve exchanges with its business partners.

A positive and significant path coefficient ($\gamma_2 = 0.28$; $p < 0.05$) confirms the second hypothesis, that is, the more strategic the role played by IT, the greater the IT management sophistication of the SME. When IT constitutes a strategic necessity or a competitive weapon, when IT is of critical importance for “core” business processes of small manufacturers, these organizations act in a coherent manner by adopting managerial practices that allow them to better manage the development and use of these technologies. These are practices such as planning, designing and evaluating IT-based applications, sustaining and favoring user participation and user appropriation of IT, preserving and developing IT resources and competencies, and seeking outside consultants to overcome internal lacks in this regard. These firms show similar coherence when they place the IT function at a high hierarchical level in the organization and render IT autonomous (with a designated manager), that is, not subordinated to the financial or accounting function as is still often the case in small business.

Due to a negative and non significant path coefficient ($\gamma_3 = -0.12$), the third hypothesis could not be confirmed. It stated that the more strategic the role played by IT, the greater the IT usage sophistication of the SME. Thus it seems that the strategic role of IT would be only indirect here, that is, through its effect on IT management sophistication. For instance, seeking internal and external integration of business processes through IT would lead the firm to better plan its use of IT and to dispose of better IT resources and competencies; only then could a more advanced technological infrastructure and applications such as ERP and e-business be implemented.

The fourth research hypothesis is confirmed by a positive and significant path coefficient ($\gamma_4 = 0.54$; $p < 0.001$), relating the firm’s IT management sophistication to its IT usage sophistication. This result increases the relevance of a strategic perspective based on IT resources and competencies, namely a resource-based view [3] to explain the level of adoption and assimilation of IT in manufacturing SMEs. Now, firms that have sufficiently developed their IT function and managerial competence and that have access to external resources are those that have adopted and assimilated the greatest number of advanced manufacturing applications, and where system quality and security are best.

Due to a non significant path coefficient ($\gamma_5 = 0.04$), the fifth hypothesis could not be confirmed. It stated that the greater the IT management sophistication, the greater the IT performance of the SME. In the absence

of a direct effect, better management of IT has nonetheless an indirect effect upon IT performance, that is, through its positive effect on the use of TI (which in turn has a direct effect on performance, as we shall see). This last result is obtained with an estimation of this indirect effect by the product of the two path coefficients ($\gamma_4 * \gamma_6 = 0.54 * 0.65 = 0.35$; $p < 0.05$).

A strong path coefficient ($\gamma_6 = 0.65$; $p < 0.001$) confirms the sixth research hypothesis, that is, the greater the small manufacturer’s IT usage sophistication, the greater the performance of its information technology. Advanced applications such as an ERP system, a transactional Web site, videoconferencing, and mobile computing, to the extent that they are effectively assimilated by SMEs, are those that are the most strategic, that is, bring the greatest “value” to these firms in the form of increased competitiveness and competitive advantage. One may recall moreover that this increased assimilation of IT is the result of better management of these technologies. In turn, this better management is the result of a more strategic vision of the role played by IT in the organization.

In total, these three factors combined explain 60% of the variance in the performance of IT. One may note here that the applications that are most affected in terms of performance are the marketing and sales applications, followed by the accounting, finance and HRM applications, and the e-business, Internet and Web applications. This last result tends to underline the more operational rather than strategic nature of the logistics, production, and distribution applications as presently implemented in the sampled manufacturing SMEs.

VI. DISCUSSION AND IMPLICATIONS

The results obtained from 44 SMEs show that IT performance is influenced in two ways. First, IT performance is directly affected by the strategic role of IT. Second, IT performance is also influenced by the indirect effect the strategic role of IT that passes first through a greater IT management sophistication, which in turn influences the IT usage sophistication, which finally contributes to IT performance.

This dual contribution of the strategic role of IT on IT performance suggests that the functional sophistication of IT alone is not sufficient to increase IT performance; it is also necessary that IT be well used by the employees. Thus, to ensure that IT fully meets its strategic role, it has to be well managed. Its development and evaluation should take into account the needs of users, involving them when conducting process analysis to make the most effective use of resources, all this being done within a structured IT

function which reflects the reality of the organization while using external resources when necessary.

The strategic role of IT has no direct influence on IT usage sophistication, however it does via IT management sophistication. It recalls that once IT is deployed and well managed, it is possible for users to enhance their strategic role. These results are in line with Westerman's [44] work on the evolution of IT. It recalls the importance of ensuring that IT should adequately support business operations, making certain that information systems work as and when they are supposed to, that their access is secure, that information is accurate, complete and correct, and that all this is done in time and within budget. Then users are able to learn and adopt the various functional applications available within the company, and to assess the quality of information they find and the links that they may develop to make better decisions.

The descriptive results indicate that for all SMEs, the benefits of IT mainly come from accounting / finance / HRM, and logistics / production / distribution applications. Then come benefits accruing from marketing / sales/ customer service applications, and to a lesser extent e-business, Internet and Web applications. This descending order of the benefits of application is consistent with the increasing complexity of IT strategic role. All companies do not cover the use of IT for electronic integration of internal and external functions, which is the most strategic role. The IT applications that are easiest to implement are often the first established, and therefore are the first to generate profits. In this study, where the benefits are cumulative by type of applications used, companies that have established several types of applications are the ones showing the highest performance from their IT. They are also those who have the most comprehensive strategic role, the more complex and more demanding.

In the context of this study, firms that gain the most benefits from their IT are those that devote the more strategic role to these technologies, manage them in a sophisticated way, and use them extensively and intensively.

VII. LIMITATIONS AND CONCLUSION

As in any empirical research, this study has some limitations that should be mentioned. Given the nature of the sample, its representativeness in relation to all SMEs limits the scope of the results. The sample firms have indeed participated in a broad diagnostic performance survey, which may reveal distinctions with the general population in terms of IT [6]. The use of perceptual measures for assessing the strategic role and performance of IT may also have induced some respondent cognitive biases, although earlier studies have also resorted to such measures [38].

Notwithstanding its limitations, this study revealed that a strategic vision of the role of IT is critical to the managerial and technological skills developed by the SME and the organizational impacts of the exploitation of these capabilities. Based on a strategic alignment perspective, however, future studies could extend the research model by examining whether the role assigned to IT depends on its fit with the SME's business strategy, structure, and environment. A formative model for measuring a most complete performance of IT such as that proposed by Gable, Sedera and Chan [16] may also be used to include, in addition to the organizational impact, individual impact, quality of IT-based systems, and quality of information produced by these systems.

REFERENCES

- [1] Andreev, P., Heatr, T., Maoz, H., and Pliskin, N. (2009), Validating formative partial least squares (PLS) models: Methodological review and empirical illustration, *Proceedings of the Thirtieth International Conference on Information Systems*, Phoenix, Arizona, pp. 1-17.
- [2] Bergeron, F., Raymond, L., and Rivard, S. (2004), Ideal patterns of strategic fit and business performance, *Information & Management*, (41:8), pp. 1003-1020.
- [3] Bharadwaj, A.S. (2000), A resource-based perspective on information technology capability and firm performance: An empirical investigation, *MIS Quarterly*, (24:1), pp. 169-196.
- [4] Brown, C.V. (1997), Redesigning the emergence of hybrid IS governance solutions: Evidence from a single case site, *Information Systems Research*, (8:1), pp. 69-94.
- [5] Caldeira, M.M. and Ward, J.M. (2003), Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises, *European Journal of Information Systems*, (12), pp. 127-141.
- [6] Cassell, C., Nadin, S., and Gray, M.O. (2001), The use and effectiveness of benchmarking in SMEs, *Benchmarking: An International Journal*, (8:3), pp. 212-222.
- [7] Chalmers, M. (1999), Comparing information access approaches», *Journal of the American Society for Information Science*, (50:12), pp. 1108-1109.
- [8] Chan, Y.E., Huff, S.L., Copeland, D.G., and Barclay, D.W. (1997), Business strategic orientation, information systems strategic orientation, and strategic alignment, *Information Systems Research*, (8:2), pp. 125-150.
- [9] Chin, W.W. (1998), Issues and opinion on Structural Equation Modeling, *MIS Quarterly*, (22:1), pp. vii-xvi.
- [10] Chwelos, P., Benbasat, I., and Dexter, A.S. (2001), Research report: Empirical test of an EDI adoption model, *Information Systems Research*, (12:3), pp. 304-321.
- [11] Cragg, P.B. (2002), Benchmarking information technology practices in small firms, *European Journal of Information Systems*, (11:4), pp. 267-282.
- [12] Croteau, A.-M. and Bergeron, F. (2001), An information technology trilogy: Business strategy, technological deployment and organizational performance, *Journal of Strategic Information Systems*, (20:2), pp. 77-99.

- [13] DeLone, W.H. and McLean, E.R. (1992), Information systems success: The quest for the dependent variable, *Information Systems Research*, (3), pp. 60-95.
- [14] Dibrell, C., Davis, P.S., and Craig, J. (2008), Fueling innovation through information technology in SMEs, *Journal of Small Business Management*, (46:2), pp. 203-218.
- [15] Dierckx, M.A.F., and Stroeken, J.H.M. (1999), Information technology and innovation in small and medium-sized enterprises, *Technological Forecasting and Social Change*, (60), pp. 149-166.
- [16] Gable, G.G., Sedera, D., and Chan, T. (2008), Re-conceptualizing information systems success: The IS-impact measurement model, *Journal of the Association for Information Systems*, (9:7), pp. 377-408.
- [17] Gefen, D., Straub, D.W., and Boudreau, M.-C. (2000), Structural equation modeling and regression: guidelines for research practice, *Communications of the AIS*, (4:7), pp.1-76.
- [18] Henderson, J.C. and Venkatraman, N. (1999), Strategic alignment: Leveraging information technology for transforming organizations, *IBM Systems Journal*, (38:2&3), pp. 472-484.
- [19] Iacovou, C.L., Benbasat, I., and Dexter, A.S. (1995), Electronic data interchange and small organizations: Adoption and impact of technology, *MIS Quarterly*, (19:4), pp. 465-485.
- [20] Jahner, S., Leimeister, J.M., Knebel, U., and Krcmar, H. (2008), A cross-cultural comparison of perceived strategic importance of RFID for CIOs in Germany and Italy, *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, pp. 1-10.
- [21] Keen, P.G.W. (1993), *Shaping the future, business design through information technology*, Harvard Business School Press, Cambridge, Massachusetts.
- [22] Kohli, R. and Grover, V. (2008), Business value of IT: An essay on expanding research directions to keep up with the times, *Journal of the Association for Information Systems*, (9:1), pp. 23-39.
- [23] Lehman, J.A. (1985), Organizational size and information systems sophistication, *Working Paper #85-18*, University of Minnesota.
- [24] MacKenzie, S.B., Podsakoff, P.M., and Jarvis, C.B. (2005), The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions, *Journal of Applied Psychology*, (90:4), pp. 710-730.
- [25] Marbert, V.A., Soni, A., and Venkataraman M.A. (2003), The impact of size on enterprise resource planning (ERP) implementation in the US manufacturing sector, *Omega*, (31), pp. 235-246.
- [26] Myers, B. L., Kappelman, L.A., and Prybutok, V.R. (1998), A comprehensive model for assessing the quality and productivity of the information systems function: Toward a theory for information systems assessment, in Garrity, E.J. and Sanders, G.L. (Eds.) *Information Systems Success Measurement*, Idea Group, Hershey, Pennsylvania, pp. 94-121.
- [27] Paré, G. and Sicotte, C. (2001), Information technology sophistication in health care: an instrument validation study among Canadian hospitals, *International Journal of Medical Informatics*, (63), pp. 205-223.
- [28] Petter, S., Straub, D., and Rai, A. (2007), Specifying formative constructs in information systems research, *MIS Quarterly*, (31:3), pp. 623-656.
- [29] Pflughoest K.A., Ramamurthy, K., Soofi, E.S., Yasai-Ardekani, M., and Zahedi, F. (2003), Multiple conceptualizations of small business Web use and benefit, *Decision Sciences*, (34:3), pp. 467-512.
- [30] Pitt, L.F., Watson, R.T., and Kavan, C.B. (1995), Service quality: a measure of information systems effectiveness, *MIS Quarterly*, (19:2), pp. 173-185.
- [31] Philip, G. and Booth M.E. (2001), A new six 'S' framework on the relationship between the role of information systems (IS) and competencies in 'IS' management, *Journal of Business Research*, (51), pp. 233-247.
- [32] Powell, T.C. and Dent-Micallef, A. (1997), Information technology as competitive advantage: The role of human, business, and technology resources, *Strategic Management Journal*, (18:5), pp. 375-405.
- [33] Premkumar, G. (2003), A meta-analysis of research on information technology implementation in small business, *Journal of Organizational Computing and Electronic Commerce*, (13:2), pp. 91-121.
- [34] Rai, A., Tang, X., Brown, P., and Keil, M. (2006), Assimilation patterns in the use of electronic procurement innovations: A cluster analysis, *Information & Management*, (43), pp. 336-349.
- [35] Raymond, L., Paré, G., and Bergeron, F. (1995), Matching information technology and organizational structure: Implications for performance, *European Journal of Information Systems*, (4:1), pp. 3- 6.
- [36] Riemenschneider, C.K. and Mykytyn, P.P. (2000), What small business executives have learned about managing information technology, *Information & Management* (37), pp. 257-269.
- [37] Seddon, P.B., Graeser, V., and Willcocks, L.P. (2002), Measuring organizational IS effectiveness: An overview and update of senior management perspectives, *DATA BASE for Advances in Information Systems*, (33:2), pp. 11-28.
- [38] Tallon, P.P., Kraemer, K.L., and Gurbaxani, V. (2000), Executives' perceptions of the business value of information technology: A process-oriented approach, *Journal of Management Information Systems*, (16:4), pp. 145-173
- [39] Tornatsky, L.G. and Fleischer, M. (1990), *The processes of technological innovation*, Lexington, Massachusetts: Lexington Books.
- [40] Uwizeyemungu, S. and Raymond, L. (2009), Exploring an alternative method of evaluating the effects of ERP: A multiple case study, *Journal of Information Technology*, (24:3), pp. 251-268.
- [41] Venkatraman, N. (1989), The concept of fit in strategy research: Toward verbal and statistical correspondence, *Academy of Management Review*, (14:3), pp. 423-444.
- [42] Venkatraman, N. (1994), IT-enabled business transformation: From automation to business scope redefinition, *Sloan Management Review*, (35:2), pp. 73-87.
- [43] Ward, J., Taylor, P., and Bond, P. (1996), Evaluation and realisation of IS/IT benefits: An empirical study of current practice, *European Journal of Information Systems*, (4:4), pp. 214-226.
- [44] Westerman, G. (2009), IT risk as a language for alignment, *MIS Quarterly Executive*, (8:3), pp. 109-121.

Developing the Mobile Service Applications of a Micropayment Platform(MPP): the Perspective of Actor-Network Theory

Jen Wel Chen

Dept. of Information Management
National Taiwan University
Dept. of Business Administration
Chinese Culture University
Taipei, Taiwan
d95725009@ntu.edu.tw

Hsiao-Chi Wu

Dept. of Information Management
National Taiwan University
Taipei, Taiwan
hciwu69@gmail.com

Ching-Cha Hsieh

Dept. of Information Management
National Taiwan University
Taipei, Taiwan
cchsieh@im.ntu.edu.tw

Abstract— Service applications for micropayment technology have become increasingly accepted among consumers. Current research on micropayment primarily discusses issues from markets' and users' perspectives, but rarely from the standpoints of enterprises'. This study examines the joint development of micropayment platforms (MPP) by various enterprises, asking: (1) why organizations develop MPP mobile service applications, (2) how they position the MPP's role, and (3) how they form project alliances. Analyzing an example of MPP development at a private Taiwanese university, this paper adopts Actor-Network Theory (ANT) concepts such as "translation" to explain not only how the different interests of stakeholders influence MPP development, but also how the identification of MPP solutions reversely conditions the selection of collaborators. Finally, this study provides practical and theoretical implications for understanding the mutually-shaping complexity and dynamics between multiple organizations and technology solutions.

Keywords: *Micropayment, mobile payment, Actor Network Theory, Near Field Communication*

I. INTRODUCTION

Service applications for micropayment technology have become increasingly accepted among consumers. In the US annually, there are now more than four hundred billion transactions under US five dollars, for a total as high as one trillion US [1]. The economic effects of these micropayments are impossible to ignore.

Micropayment platforms (MPP) are a standard inter-organization information system, requiring the cooperation of various enterprises for their development and implementation. They are subject to the interests and intentions of the partnering institutions, whose preferences are also conditioned by the group-selected MPP technological solution. As a result, the MPP development process is both highly complex and dynamic.

Current research on MPP primarily emphasizes technological development[2, 3], consumer equality[4], and the connection schemes of various payment systems[5]. Research on MPP is diverse however, and other studies examine the use of electronic purse systems in different countries[6, 7], offer economic analysis of micropayment systems[8, 9], evaluate transaction costs[10, 11], explain why MPP succeeds and fails[12-14], and identify new mobile payment applications [8, 16, 17].

All of this research however, starts from the perspective of markets and users, and thus neglects the standpoint of enterprises. In response, this study discusses the mechanisms through which stakeholder participants in this technology form partnerships, how their respective interests interrelate, and how these interests influence the infrastructure and functions of MPP. These issues are examined through a special case study involving a three-year inter-organization MPP that employs Near Field Communication (NFC) to develop mobile service applications for a private Taiwanese university. Specifically, this study answers the following questions: (1) why the enterprise developed MPP mobile service applications, (2) how it positioned the MPP's role, and (3) how it formed alliances for the project; i.e., how it recruited and negotiated with other organizations to jointly develop MPP.

Adopting Actor-Network Theory (ANT), we analyze the project formation and choices of NFC and MPP in order to understand why this university developed MPP techniques for integrated campus services.

II. MICROPAYMENTS

A micropayment usually refers to a single transaction of less than US five dollars [1]. The proportion of transaction fees for micropayments is also much higher than other payment channels like credit cards or checks. These transaction costs are critical for the acceptance of micropayments[10], although the following two features transform micropayments into a real payment tool: (1) it serves as a unit of accounting or a standard of value, (2) and it can be used as an interchangeable medium or payment tool to facilitate transactions and reduce redundant transaction costs[15]. The current transaction fee for micropayment is proportionally high, but the electronic payment system is able to provide the advantages of lower transaction costs and quicker transaction times. Micropayments would be more likely to succeed through the linkage of electronic payment system.

Current techniques for electronic payment processing can be divided into three types:(1) credit-based solutions, (2) account-based solutions, and (3) value-stored application[1]. These payment solutions comfortably fit into the micropayment environment due to their cost, speed and convenience. Micropayments are smoothly developed into electronic payment systems through these three channels.

In mobile commerce, micropayment is also an important and interesting service for mobile financial applications. For example, a payer can employ a mobile device through a local wireless network to purchase desired goods or services. Micropayments can also assist with telephone calls to numbers where per-minute charges equal the cost of vending items, or with pre-paid purchases from a service provider, bank, or credit-card company[16]. Dahlberg et al. [17] note that the majority of research on these applications for mobile payment focuses on technological development and consumer acceptance of technology; few consider the perspective of payment service providers.

Within the current business environment, information technology (IT) is a key tool for organizations to promote their competitive advantages and differentiate themselves from their competitors[18, 19]. The development of MPP however, requires collaboration between various partners. The intentions and interests of these involved organizations increase the complexities of the MPP. Furthermore, the social relationships of these organizations affect this development process, as does the essential nature of MPP technology. These factors prevent the evaluation of MPP development by simply looking at the strategies and intentions of the involved parties.

III. ACTOR-NETWORK THEORY

In IT research fields, Actor-Network Theory (ANT) is a widely accepted approach for understanding the complex social interaction of technological changes[20] and explaining the socialized processes of technology development and introduction in different contexts[21]. ANT was developed within the field of sociology of science and technology[22, 23]. ANT scholars thought that researchers should observe the fusion and relationships among science, technology, and society by tracing science in action, i.e., boundlessly following the actions of actors[22]. A key feature of ANT is that actors can be both human and nonhuman, e.g., technological artifacts[20].

The purpose of ANT is to examine, through the actors in a network, the actions and motivations of actors and the heterogeneous network that links relationships and aligns interests. ANT can explain how actors (human and technological) with different interests jointly create a relatively stable arrangement of technology[23], so as to express and understand the features and approaches of information technology development. One notable example of this approach is Michel Callon [23], who utilized the word “translation” to refer to the creation of actor-networks. Here, a focal actor is the key actor driving the process of recruiting other actors into the actor-network, transferring its intention, purpose, recognition and behavior to them so that they share common behavior and viewpoints. In this way, an actor-network is not just a simple combination of actors, but rather the seed of a focal actor, who redefines and rearranges the interests, roles, functions and positions of each partner into a new actor-network. The process of translation consists of four stages:

- **Problematization:** During the problem-formulating stage, the focal actor defines the identification and

interests of the other actors who share initial common interests. By defining problems and solutions, concerned actors can be confirmed and their roles identified.

- **Interessement:** The focal actor convinces the other actors that the interests defined by the focal actor are actually in line with their own interests.
- **Enrollment:** The focal actor persuades the other actors to accept a set of strategies and the definition of their roles in the developing actor-network.
- **Mobilization:** The focal actor, by using a set of methods, ensures that the actor-network operates according to mutual agreement so the network remains stable.

During the network building process, the focal actor aligns the interests of all actors according to the network’s interest and recruits recognized actors to establish the actor-network. Accordingly, the focal actor transforms itself into an obligatory passage point (OPP) for all network actors. This process of translation is a suitable model for understanding the interaction of multiple social groups in the development of MPP.

IV. RESEARCH METHOD

This study employs the case study research method. Case studies allow the researcher to investigate the interrelationships and dynamics of research phenomena and contexts under a natural setting[24]. MPP involves various organizations whose actions and interests influence and are influenced by MPP development. Thus, MPP development is not strictly predetermined, but rather emerges from continuous party alignments and negotiations.

Data were collected from March, 2006 to December, 2008 through participant observation, semi-structured interviews, meeting minutes, and project files, as depicted in Table I. These multiple sources helped ensure data authenticity and validity.

TABLE I. THE TYPES AND ILLUSTRATIONS OF EMPIRICAL DATA

Data type	Illustrations
Participant observation	Observation period: 2006.3-2008.12 One of the authors works at CU, he participated in digital campus projects
Semi-structured interviews	Eight single person interviews. Each interview lasted 90 to 120 minutes. Some interviewees were interviewed twice depending on the situation.
Meeting minutes	99 recorded minutes, including discussions on the cooperation and negotiation of enterprises, technology, managers’ meetings, and technology group meetings.
Documents	104 files , including project reports, proposals, technical documents, memorandums, official documents, presentations, and historical data.

V. CASE DESCRIPTION

CU (anonymous) was the first university to issue student ID smartcards (UPass) in Taiwan. CU invests significant resources on IT and aims to develop its digital campus as an important characteristic and competing advantage. In 1998, CU launched the contact IC card as a student ID smartcard, and gradually integrated various campus services. In 2005,

CU converted UPass into a RFID (radio-frequency identification) stored value card, and integrated twelve campus services in three major fields: building access control, campus administration services, and payment tools. UPass campus services develop in three periods:

A. The period of digital campus presentation

In April, 2006, CU held a press conference to demonstrate the further development of its digital campus vision with a joint program between a transportation company(EZCard), local bank(Bank C), carbonated beverages vendor(Vendor C), international software company(SWHouse-1), and a system installation hardware company(HWHouse-1). More than 40 schools, several high technology companies, and related government organizations participated in the press conference mentioned above. Afterwards, several schools showed their own interest in such programs, and asked CU and Bank C to help promote the MPP of integrated campus services for their own schools. The development team of UPass thus began considering the possibility of promotion.

B. The period of MPP development

After the conference, CU discussed its goals for the next stage of campus services and its new technology. The school regarded MPP as the key development for future campus services, and decided that future expansion of MPP was best handled by other institutions such as EZCard rather than by the university itself.

CU tried to convince Bank C to develop certain MPP (as Solution-2) suitable for its campus. These efforts were unsuccessful however, as the bank’s existing MPP features (QPay, as Solution-3) were credit card based, which by law face certain restrictions on school campuses. Bank C was required to adjust its QPay mechanism and supply related equipment to interested university campuses—measures the Bank eventually abandoned despite university demand, due to legal concerns regarding their credit card business.

Meanwhile, EZCard successfully issued student ID smartcards for CU as part of value-stored application (Solution-7) jointly supported by Bank C and three other local banks. Here however, legal concerns still existed, as transportation passes by law may not provide MPP functions.

Despite these setbacks, CU continued to seek new MPP alternatives. It found that the technology scheme of online MPP(as Solution-4) issued by local Bank S was quite similar to current MPP (Solution-1). CU recruited Bank S as a cooperating partner to help expand MPP. Coincidentally, at the time a leading retail enterprise (Retail K) also intended to seek the cooperation of a bank to promote its own MPP (as Solution-5). As a result, Bank S ceased negotiations with CU in order to actively focus on courting Retail K—a battle it eventually lost to Bank C.

CU’s latest MPP partnering solution involved SmartPay, an account-based system (Solution-8) promoted by a bank association (Org-2). SmartPay serves as a account based MPP system in collaboration with 22 banks. Its value can be stored through ATM transfers, which is convenient for many

campus applications. Currently however, Solution-8 is still under incubation, and not yet officially released.

It should be noted that during its attempts to expand MPP services, the university’s existing MPP system was still operating smoothly without urgent need for replacement. With the success of its press conference and subsequent additional interactions with other enterprises and universities, CU simply expected new possible projects to promote campus services.

C. The emergence of a NFC solution

Through SWHouse-1, CU discovered the installation of a NFC campus application solution at a university (AU) in Austria. The IT department dean of CU visited AU in 2007, and CU determined that the techniques of NFC were compatible with the existing campus system. Encouraged by SWHouse-1, CU formed an NFC mobile network and found a new banking partner, Bank U. After explaining its existing campus services as well as the consumer finance opportunities combined in NFC cellular phones, Bank U was deeply motivated and willing to offer its cooperation.

With the MPP-NFC solution (Solution-9) as its target, CU and Bank U applied for a government technology project encouraging mobile payments as a financial innovation (PMPTFI). Bank U also recruited a systems integration company, SWHouse-2, to jointly develop NFC micropayment and clearing mechanisms for CU’s campus. Still in need of a telecommunications partner, CU demonstrated its campus services to Taiwan’s leading telecom firm, Mobile K, which was willing to participate in the project. CU then invited a professor from AU as a consultant, and the NFC mobile payment platform was officially established.

Overall, the 2006 press conference of the CU student ID smartcard acted as a catalyst, leading to the involvement and competition of as many as six banks, four universities, and nine possible MPP solutions. Four groups were formed during the process until finally, the governmental technology project PMPTFI was established. Figure 1 and Table II display the thirty actors and nine MPP solutions involved during this MPP development process.

TABLE II. CU MPP SOLUTIONS

MPP	Focal actor	Features
1	CU	Account-based, on-line and off-line dollars by CU
2	CU	Account-based, on-line and off-line dollars by Bank C
3	Bank C	QPay, credit card based, off-line dollars
4	Bank S	Ecoin, account-based, on-line dollars
5	Retail K	ICash, value-stored application, off-line dollars
6	Bank C	QPay II, top-up solution, off-line dollars
7	EZCard	EazyCard, transportation ticket, value-stored application, off-line dollars. Since April 2010, used for payment at over 10,000 locations in Taiwan.
8	Org-2	SmartPay, account-based, on-line dollars
9	Bank U Mobile K	Mobile NFC, value-stored application, on-line and off-line dollars

VI. ANALYSIS

From the point when CU successfully demonstrated the integration of campus services with student ID smartcards, the process of seeking the next IT development solution can be divided into three periods. Here, we demonstrate the changes in relationships between CU and other actors through the employment of ANT.

This study employs ANT as a theoretical lens, treating the various MPP techniques as equal actors. Using ANT, we can clearly understand the detailed process and dynamics of the RFID solution to the NFC mobile payment project. Adopting the analysis of the four elements of translation proposed by Callon[23], we discuss the phenomena from the following three aspects.

A. Problematization

After the press conference, CU defined the problem as “innovation diffusion” to attract the participation of other actors. However, other actors did not proactively seek a partnership. This forced CU to redefine the problem in order to find other actors. The recruitment of Bank S is a typical example. CU redefined the problem as “MPP development in campus” as a way to encourage and engage banks. Thus, before the network successfully formed, CU had to redefine the problem repeatedly in order to acquire partners. Table III presents the ways CU defined its problem at different periods.

TABLE III. PROBLEMATIZATIONS FOR CU

Time period	Problematization
Digital campus presentation	“Innovation diffusion” enables integration between payment card and campus services
MPP development	“MPP development in campus” implies the increasing card volume for banks
Emergence of NFC solution	Technology innovation of campus services keeps CU leading IT in Taiwan

B. Interesement and Enrollment

Under different contexts, CU defined different questions in order to attract different actors in various domains. Even the actors within the original network, e.g. Bank C, still sought other opportunities for linking other actors to form new network relationships. Issuing student ID smartcards enabled banks to increase their card volume, although it did not convince them of further participation. CU did not have alternative strategies to attract banks for developing MPP. Furthermore, an interesting new target (the MPP of Retail K) opened for banks in the MPP market. These consequences led to multiple actors and an unstable network relationship in the period of MPP development. This situation was not resolved until CU discovered its NFC mobile solution and recruited partners in Bank U and SWHouse-2.

C. Mobilization

Bank U was enrolled into the network owing to the interests defined by CU. SWHouse-2 was recruited as well. These recruited actors formed a network through mobilization through the alignment of their interests. Meanwhile, MPP solutions act as an equal actor to filter and

determine the formation of the actor-network. For example, Solution-9 led to the enrollment of telecommunication enterprise Mobile K, whereas Solution-3 and Solution-7 were unable to form stable networks due to legal constraints.

However, actors may also betray their original network in favor of the appeal of other networks, or simply cease to invest resources into the original network. For example, Bank C escaped from their original network through its attraction to Solution-5 with Retail K. This eventually led to CU’s association with Bank U, SWHouse-2, and Mobil K in the application of their governmental project, i.e. PMPTFI, which becomes the OPP in order to form a stable relationship of network. This action may prevent the betrayal of other actors, help negotiate a precise goal, gain economic support, and direct all the interests of actors toward a common goal. When more resources are invested, this will facilitate the participation of other actors as needed.

In ANT, CU MPP development is also a process of “translation”. The actor’s interests are firstly translated into specific needs which will be further translated into the system solution. The system solution will then be adopted by actors, who translate it into the context of their specific work practices. Thus, CU recruits different actors for each specific technology solution. After CU recruits actors, i.e. Bank U, SWHouse-2, and Mobile K, into the actor-network in the last stage, the technology scheme has to be changed in order to include their requests. The change of technology scheme stimulates the generation of mobile NFC solution which turns into the core of MPP application developments.

VII. DISCUSSION

This study analyzes the development process of the MPP NFC solution, elaborating on the complex relationship between multiple organizations and technology. The case study above demonstrates how CU MPP development differs from the conventional development of information systems. The processes of MPP development are emergent and unexpected, support Orlikowski’s [25] arguments, and are gradually clarified through the contact of various actors. Each actor pursues its own interest as well as takes into account the common interest of other actors. Through interactions among actors, the cooperative relationship for inter-organization is formed. Even under this cooperation however, actors still seek other opportunities for linking other actors to form more beneficial cooperative partnerships. CU must continuously recruit other organizations, revise its interests and its micropayment solution, and prevent actor betrayal. Therefore, the goal of technology development, the scheme of the MPP system, and the collaborative groupings are not pre-planned but rather emergent and unexpected. Through this process, a leading organization may form an inter-organization cooperation network.

From the ANT perspective, MMP technology will have some effects on selection and formation network. Therefore, these effects trigger interaction among network actors. The selection of network also limited technology choices. Base on history of developing NFC platform and the role of participants, that can be understand more clearly the nature

of this case. This study realizes that CU continuously repositioned the direction of MPP development over a period of three years. The complexities and dynamics presented in the case can be summarized as follows.

A. MPP development as a process of redefining roles and interests

Requiring multiple partners in the development of MPP usually detracts from the original cooperative theme on account of opposing thoughts and interests. The CU MPP position shifted from original innovation diffusion, to the development of MPP, then to the NFC solution. This demonstrates that problematization is not a one-time activity but a continuous ongoing process [26]. We found that CU not only continued discussing possible developments of campus services and innovation of their MPP scheme, but also sought the cooperation of other actors and redefined their actor-network. Thus, the IT developmental processes of CU campus services are a mixture of changes and eliminations. This process is based on a multi-directional model, rather than a linear model. Technology can have more than one developmental result.

B. Each actor as a latent actor-network

CU actively pursued cooperative organizations to develop its campus services, but the final MPP solution network was promoted by outside actor SWHouse-1. It forced CU to reposition the MPP, i.e. consumers' services on NFC cellular phones. Moreover, the request by several universities to Bank C for MPP made Bank C reconsider the Solution-2 proposed by CU. To adapt Solution-2, Bank C had to upgrade QPay (Solution-3) to QPay II (Solution-6) for campus needs.

Past studies of ANT discuss the development of information systems focusing on the actor-network created by focal actors [26, 27]. However, Monteiro[29] suggests actors may be included into a new, more complete actor-network. In other words, other actors in the network can also play the role of focal actor. According to their own interests, they will seek other outside opportunities or link other actors to form a new actor-network. This latent actor-network influences the interests and stability of the original network.

This implies that to understand the changing processes of MPP development, researchers must not only observe the network of CU recruitment, but also the networks of other actors. We find that from a macro level (based on micropayment markets), each actor can seek other resources to form other latent actor-networks that may influence and break the existing stable network relationship, and result in changes to the MPP solutions.

C. Temporary actors as the catalyst of actions

Once CU can not align mutual actors' interest, the network will not be formed and the developments of MPP development will not be carried out. In previous research on ANT, temporary actors in unformed network are usually neglected. Groothuis and Akbar[28] argue that temporary actors act as a catalyst, and they can affect actor decisions and actions within the network. In case of CU, these

temporary actors, though not being part actor of the final network, did have critical influences and trigger further actions by CU.

For example, when School-2 consults with CU about the development of campus smartcards, it stimulates CU to actively think how to diffuse its results to other campuses. Furthermore, with the contact of the NFC forum, CU is inspired to realize the NFC technique, and finally promote the NFC solution of MPP. Retail K's issuance of its ICash card quickly disrupts the negotiation and cooperation between CU and Bank S. ANT scholars should consider how to account for these influential temporary actors.

VIII. CONCLUSION

This study applies the ANT perspective to the development of an inter-organization system. We examine the complex relationships between organizations and technology under the research setting of the development of MPP in CU. The characteristics of the technology solution filter the choices of participant organizations when an enterprise chooses that solution. Meanwhile, when recruiting the participant organizations, the alignment of mutual interests forces enterprises to continuously adapt the goals of MPP and adjust the technology solution before the goal of campus services can be set.

Current research on micropayment primarily discusses issues from markets and user perspectives, but rarely from the standpoints of enterprises. The main contributions of this study are two fold. (1) In practice, the development of an inter-organization information system, e.g. MPP, is usually an emergent and unexpected process. The organization must continuously redefine its own role and technology scheme, and also align mutual actors' interests when searching for collaborative organizations. From an ANT perspective, our study finds that problematization is not a one-time activity but an on-going process. (2) Theoretically, there are rare phenomena in previous ANT literature. One is that each actor of a network represents another latent actor-network, which affects and breaks existing stable network relationships. The other is that temporary actors do influence the focal actor to redefine the problems and roles of other actors.

Our findings provide two main directions for future research. Additional effort is needed to investigate the degree of connectivity among various actor-networks. Furthermore, given the results above, more information is needed on how temporary actors influence changes in actor-networks.

REFERENCES

- [1] S. Ching, A. Tai, J. Pong *et al.*, "Don't Let Micropayments Penalize You-- Experience From The City University Of Hong Kong," *The Journal of Academic Librarianship*, vol. 35, no. 1, pp. 86-97, 2009.
- [2] S. Jarecki, and A. Odlyzko, "An efficient micropayment system based on probabilistic polling," in *Proceedings of Financial Cryptography Conference, Lecture Notes in Computer Science*, Springer Verlag, 1997, pp. 173-191.
- [3] R. Rivest, and A. Shamir, "PayWord and MicroMint: Two simple micropayment schemes," in *4th International Security Protocols Conference*, 1996, pp. 69-87.

[4] S. Yen, "PayFair: A prepaid Internet micropayment scheme ensuring customer fairness," *IEE Proceedings-Computers and Digital Techniques*, vol. 148, pp. 207, 2001.

[5] R. Parhonyi, D. Quartel, A. Pras *et al.*, "An interconnection architecture for micropayment systems," in Proceedings of the 7th international conference on Electronic commerce, Xi'an, China, 2005, pp. 640.

[6] F. Prior, and J. Santoma, "The Use of Prepaid Cards for Banking the Poor : Comparative Study Analysing the Development of Prepaid Systems in the United States and Europe," *IESE Research Papers*, 2008.

[7] L. Van Hove, "Electronic purses in Euroland: why do penetration and usage rates differ?," *SUERF Studies*, M. Balling, ed.: Vienna, Austria: Société Universitaire Européenne de Recherches Financières, 2004.

[8] Y. Au, and R. Kauffman, "The economics of mobile payments: Understanding stakeholder issues for an emerging financial technology application," *Electronic Commerce Research and Applications*, vol. 7, no. 2, pp. 141-164, 2008.

[9] C. Schmidt, and R. Muller, "A framework for micropayment evaluation," *Netnomics*, vol. 1, no. 2, pp. 187-200, 1999.

[10] B. Neuman, "Security, payment, and privacy for network commerce," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 8, pp. 1523-1531, 1995.

[11] I. Papaefstathiou, C. Manifavas, and B. Capital, "Evaluation of micropayment transaction costs," *Journal of Electronic Commerce Research*, vol. 5, no. 2, pp. 99-113, 2004.

[12] W. Tan, and S. Chen, "An analysis of the factors influencing success of bank-issued micropayment system in Taiwan," *Journal of Systems and Information Technology*, vol. 10, no. 1, pp. 5-21, 2008.

[13] W. Tan, "Factors Affecting Non Bank-Issued POS E-micropayment Choice: A Study of Taiwan Market," *Journal of Internet Banking and Commerce*, vol. 13, no. 3, <http://www.arraydev.com/commerce/jibc/2008-12/TAN.pdf>, 2008.

[14] C. Clark, "Shopping Without Cash: The Emergence of the E-purse," *Economic Perspectives*, *Federal Reserve Bank of Chicago*, issue Q IV, pp. 34-51, 2005.

[15] R. Weintraub, *Introduction to Monetary Economics: Money, Banking, and Economic Activity*: Ronald Press Co., 1969.

[16] U. Varshney, and R. Vetter, "Mobile commerce: framework, applications and networking support," *Mobile Networks and Applications*, vol. 7, no. 3, pp. 185-198, 2002.

[17] T. Dahlberg, N. Mallat, J. Ondrus *et al.*, "Past, present and future of mobile payments research: A literature review," *Electronic Commerce Research and Applications*, vol. 7, no. 2, pp. 165-181, 2008.

[18] B. Ives, and G. Learmonth, "The information system as a competitive weapon," *Communications of the ACM*, vol. 27, no. 12, pp. 1193 - 1201, 1984.

[19] C. Wiseman, and I. MacMillan, "Creating competitive weapons from information systems," *Journal of Business Strategy*, vol. 5, no. 2, pp. 42-49, 1984.

[20] G. Walsham, "Actor-network theory and IS research: current status and future prospects," *Lee, AS, Liebenau, J. and DeGross, JI (Eds.) Information Systems and Qualitative Research*, pp. 466-480, 1997.

[21] M. Akrich, "The de-scription of technical objects," *Bijker, W. and Law, J. (Eds.) Shaping technology, building society: Studies in Sociotechnical Change*. Mass, MIT Press, pp. 205-224, 1992.

[22] B. Latour, *Science in action: How to follow scientists and engineers through society*: Harvard Univ Pr, 1987.

[23] M. Callon, "Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay," *John Law (Eds) Power, Action and Belief: A New Sociology of Knowledge*, vol. London: Routledge & Kegan Paul, pp. 196-233, 1986.

[24] R. Yin, *Case study research*: Sage publications Newbury Park, Calif, 1994.

[25] W. Orlikowski, "Improvising Organizational Transformation Over Time: A Situated Change Perspective," *Information Systems Research*, vol. 7, no. 1, pp. 63-92, 1996.

[26] S. Sarker, and A. Sidorova, "Understanding business process change failure: An actor-network perspective," *Journal of Management Information Systems*, vol. 23, no. 1, pp. 51-86, 2006.

[27] B. Ajad, and S. Faraj, "Examining Alignment of Frames Using Actor-Network Constructs: The Implementation of an IT Project," *AMCIS 2007 Proceedings*, pp. 258, 2007.

[28] A. Groothuis, and Y. Akbar, "Organizational transformation—from multinational to global: An early dynamic modeling perspective," *Global Business and Organizational Excellence*, vol. 26, no. 4, pp. 47-61, 2007.

[29] E. Monteiro, "Actor-Network Theory and Information Infrastructure," *From Control to Drift. The Dynamics of Corporate Information Infrastructures*, C. C. Associates, ed., pp. 71-83, Oxford: Oxford University Press, 2000.

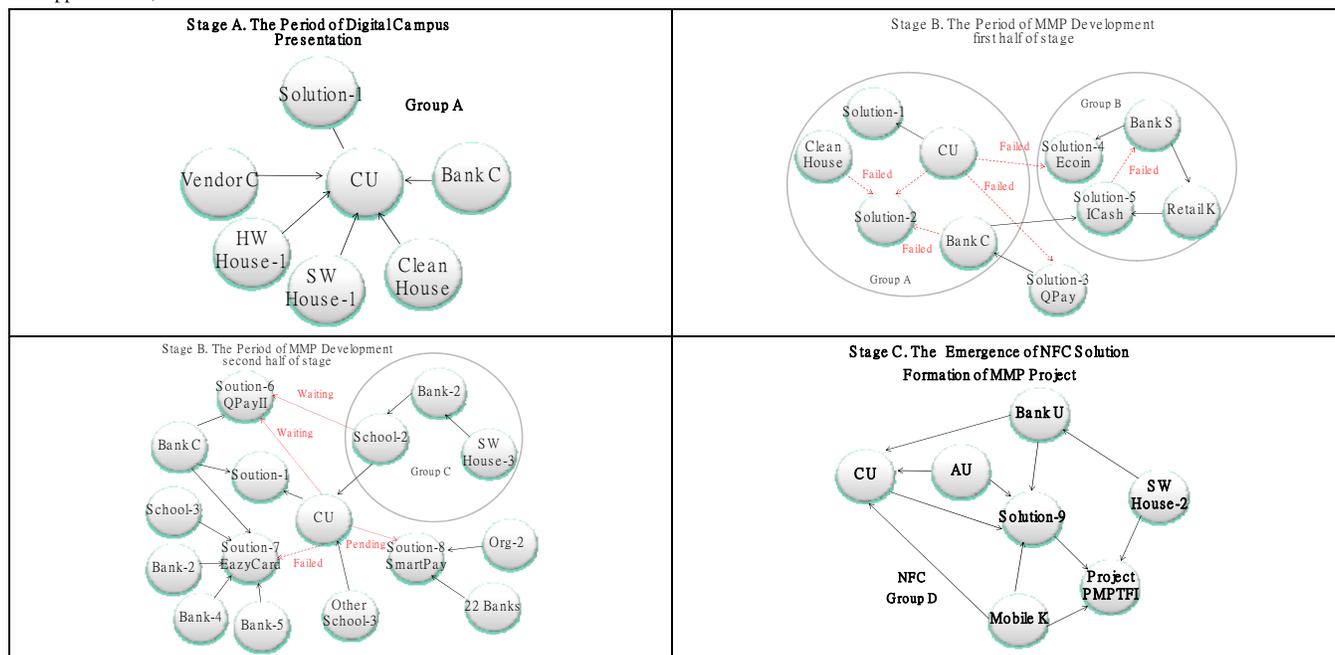


Figure 1. Change process of developing the NFC solution of MPP in CU

Decision Method of Training Data for Web Prefetching

Zhijie Ban Feilong Bao

College of Computer Science, Inner Mongolia University, China

e-mail: banzhijie@imu.edu.cn csfeilong@imu.edu.cn

Abstract—Web prefetching is an effective technique to reduce user-perceived latency. Most studies mainly focus on prediction algorithm but they ignore selection strategy of training data which is an important part of web prefetching. This paper presents a decision method based on monitoring prediction precision. It divides user access sequence into different data blocks and the changing features of prediction precision among data blocks indicate whether some training data is outdated. According to the varying trend of prediction precision, some user access requests are inserted into or deleted from training data. We use two real web logs to examine this proposed method and the simulation shows that our method can significantly improve prefetching performance.

Keywords—web prefetching; sliding window; training data

I. INTRODUCTION

Web prefetching technique is one of the primary solutions used to reduce user-perceived latency. The spatial locality shown by user accesses makes it possible to predict future accesses from the previous ones [1][2]. Web prefetching system makes use of these predictions to preprocess user requests before they are actually demanded. Part of the network latency can be hidden if prefetching system prefetches those pages which are very likely to be demanded in subsequent requests.

To predict the user's next request, a number of prediction approaches were presented, which had achieved an acceptable performance [3]. In the web prefetching technique, part of user access sequence is used as training samples to construct prediction model before user requests are predicted. By training with samples, prediction model includes user access patterns and some important information, which provides a foundation for predicting the user's next request page. Thus training data is very important to correctly predict user requests. However, few studies focus on decision method of training data. Many researchers random select one part of user access sequence as training samples and another part is used as test samples. Nanopoulos et al. used 75 percent of a week Clarknet log available from the site <http://ita.ee.lbl.gov/html/traces.html> as training data and 25 percent as test data [4]. Sarukkai presented that 40000 samples of the EPA-1995 server log were used as training samples and the remaining as test samples [5]. Shi and Gu used 80 percent of one month's NASA -1995 log to train prediction model and 20 percent

as test data [6]. Only the papers slightly talked about training data problem [7][8][9]. In order to verify client-based web prefetching experiments, Lan and Ng [7] obtained a proxy trace whose web pages were requested by different users. Then, the log was partitioned into a number of the single user's access sequences. Finally, they randomly selected continuous 14 days web accesses from every user's log to train prediction model and the fifteenth day's user requests were predicted according to the constructed prediction model. During the experimental period, they found that the web accesses of 14 days were enough for describing user access patterns. So two-week log was selected for every user as training samples. In order to examine the web prefetching performance, Davison shown the prediction model was not trained before predicting the next user request [8]. He considered that this method was better near to the real network environment. But the prediction precision is very low if the prediction model is seldom trained in the real prefetching system. In the low precision's condition, network resources such as network bandwidth are wasted if predicted pages are prefetched. Domènech and Sahuquillo studied how training data to influence prefetching performance with two different prediction models and 4 different logs [9]. They compared prefetching performance using the old and current web log, but they did not study how to decide training data.

This paper presents one decision approach of training data based on our previous work [10]. It partitions the user access sequence into different continuous data blocks according to the access time of every request. Based on the changing trend of prediction precision among different data blocks, our method decides whether web accesses are deleted from or added into training data. As a result, prediction model space is decreased and prefetching performance is improved.

The rest of the paper is organized as follows. Section 2 presents the related background. Section 3 describes the decision strategy of training data and its algorithm. Section 4 gives the details of our experiments and testing results. Section 5 is the summary and conclusions.

II. RELATED WORK

There is an important set of research works concentrating on prefetching techniques to reduce the user perceived latency. Various prediction models have been proposed to model and predict a user's browsing behavior on the web. Markatos and Chronaki proposed a Top-10 approach which combined the popular documents of the servers with client

Supported by the program of higher-level talents of Inner Mongolia University (Z20090137) and national innovation experiment program for university student (101012623)

access characteristics [11]. Web servers regularly pushed the most popular documents to web proxies, and then proxies pushed those documents to the active clients. But the approach only made use of page access frequency. In order to solve the problem, the study in [12] presented a prefetching algorithm based on prefetching in the context of file systems [13]. The server built a dependency graph (DG) where an arc from node A to B meant that B was likely to be accessed within a short interval after an access to A. Each arc was labeled the conditional probability. But the DG model was not very accurate in predicting the user browsing behavior because it only considered first-Order dependency [4] and did not look far into the past to correctly discriminate the different observed patterns. Thus, the studies in [14][15] described the use of a kth-Order Markov model for user request patterns. In a kth-Order Markov model, each state represented the sequence of k previous requests, and had conditional probabilities to each of the next possible states. However, it is likely that there will be instances in which the current context is not found in a kth-Order Markov model if the context is shorter than the order of the model. Therefore, the PPM (Prediction by Partial Matching) model [16][17][18] which originates in the data compression community, overcomes the problem. It trained varying order Markov models and used all of them during the prediction phase. Fan et al. studied how user access latency could be reduced for low-bandwidth users by using compression and PPM prediction model between clients and proxies [17]. Bouras et al. studied prefetching's potential in the Wide Area by employing two prediction models [19]. These PPM models do not implement the online update and timely reflect the changing user request patterns. An online PPM with dynamic updating is presented [20]. But most of them arbitrarily take a part of web log as training set and another part as prediction set. Only the studies in [7][8][9] slightly mention the training data problem. Lan and Ng proposed a client-based web prefetching management system, which was based on the caching schema of Netscape Navigator [7]. In the experiments, users submitted their web access requests through their own machines to the proxy server, and their prefetching system obtained each log file that contained the log of each individual user's web access requests within a 2-week consecutive time period. A 2-week time period was chosen because it was sufficient to show the web access pattern of each individual user based on their observations during the experimental period. Thus, they randomly chosen a 2-week consecutive time period for each user to represent the access history of the user as long as the user accessed the web on the fifteenth day, the day after the 2-week consecutive time period. But Domènech and Sahuquillo considered that the length of training period may impact on prefetching performance, either improving or degrading it [9]. In addition, this length may involve a high amount of information and therefore important computer resources are consumed. Thus, they analyzed that how the training affects the prediction performance using

current and old web traces. Their experimental results showed that while in old traces the training, in general, improves performance, when using recent traces this training may degrade performance because users' access pattern had changed. Davison evaluated prediction algorithms without previous training [8]. This procedure was argued to be more realistic than freezing the learning after a training period [9]. But all of them do not study how to dynamically determine training data according to different user access behaviors.

III. DECISION APPROACH

In this section, we specify concept definition, and give decision strategy and decision algorithm.

A. Concept definition

We firstly give some related concepts before decision method is introduced.

Definition 1 *User access sequence* is an orderly sequence composed of a series of two-tuples such as $\langle T_1, I_1 \rangle, \langle T_2, I_2 \rangle, \langle T_3, I_3 \rangle, \dots$, where $T_i (i=1,2,3,\dots)$ is the access time, I_i denotes the entity, T_j is larger than T_i if j is larger than i .

The time of two-tuples has strong restriction and denotes the absolute time of user request. The entity of two-tuples represents every request's attributes. Suppose the entity I includes k attributes $\{X_0, X_1, \dots, X_{k-1}\}$, where the value range of the attribute X_i is $d(X_i)$, the attribute space of the entity I is $\{d(X_0), d(X_1), \dots, d(X_{k-1})\}$. In the server's log, every $\langle T, I \rangle$ corresponds to one user request record, where T represents the user absolute request time and I mainly includes IP address, the request page's URL and so on.

Definition 2 *Sliding window* is defined a user access sequence including h user requests, where h is the number of user requests in the sliding window.

Figure 1 gives a sliding window's sketch map with h user requests. In order to describe simple, the two-tuples of user request sequence is denoted as a_j , where j is the relative access time. In the sliding window, a_i is the eldest user request, a_{i+h-1} is the newest one and a_{i+h} is the user request which will slide into the sliding window.

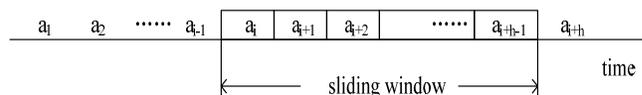


Figure 1. Sliding window with h user requests

Definition 3 *Data block* refers to one user request sequence and all requests are ranked according to the access time from the eldest one to the newest one. Partition of data blocks may take time segment or request number as dimension. We choose the former because there may exist a large number of requests in a short time. When the emergent event happens, data block using fixed request number as dimension can not represent user access behaviors while

data block with the time dimension better reflects user access features.

Definition 4 Window includes one user access sequence during a period of time and is partitioned into n data blocks according to the same time dimension. The label of data blocks in one window varies from 0 to $n-1$. Figure 2 depicts a window's sketch map with n data blocks. In the window, the data block labeled 0 is the eldest and one labeled $n-1$ is the newest. The user request number of every data block may be different while the time periods are the same.

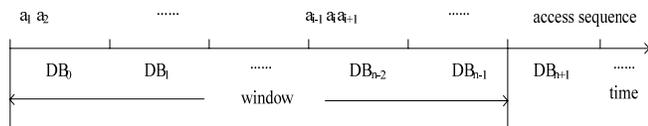


Figure 2. Window with n data blocks

B. Decision strategy

The right training data is important for constructing prediction model and predicting the user next request page. If training data includes too little access requests, the relevant user requests may be forgotten so that some correctly access features may be deleted. If training data includes excessive user accesses, the prediction model do not also represent the browsing characteristics of current users because it may include some outdated user access patterns and browsing information.

We use one sliding window SW and two windows (W_S and W_L) to dynamically adjust training data to reduce the prediction model's space and improve prefetching performance. The sliding window SW includes the total user access sequence in the prediction model. W_S is called the small window which includes some continuous data blocks. W_L is called the large window which includes W_S and other some continuous data blocks. W_S is a part of the large window. The sliding window SW includes W_L and the newest user access requests which can not compose one data block. In order to decide training data, the large window size is adjusted according to prediction precision's changing features among data blocks of the small window so that the sizes of the sliding window and W_S change.

Figure 3 gives the relation between the small window and the large window. In Figure 3, the total user access sequence is regarded as a series of user requests. It is denoted a_1, a_2, a_3, \dots , where a_i stand for user request and i is the relative access time of the i th user request. The user access sequence is partitioned into some data blocks according to the same time, where DB_0 is the eldest data block and DB_n is the newest one in the large window. The large window W_L includes $n+1$ continuous data blocks and the small window W_S includes m continuous data blocks, where m is smaller than n , and the m data blocks are the newest in the large window.

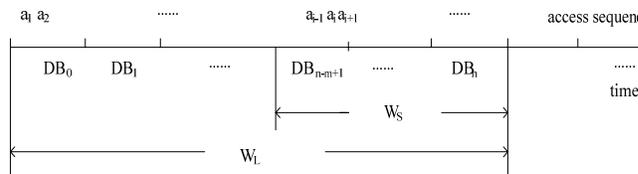


Figure 3. Relation between small window and large window

In order to specify the relation between data blocks of the large window and the user access sequence used to train prediction model, Figure 4 gives the relations among the large window, the small window and the sliding window. By the time dimension, the total user access sequence is partitioned into some continuous data blocks and some subsequent user requests which can not form one data block. The sliding window represents the total user access sequence which is used to construct prediction model. The large window includes all of data blocks labeled from 0 to n . The small window is a part of the large window, whose data blocks are labeled from $n-m+1$ to n .

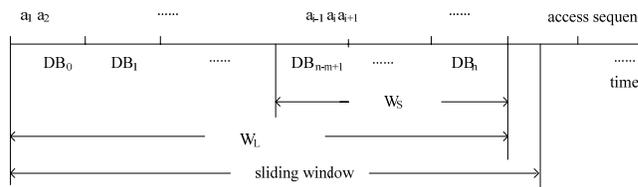


Figure 4. Relations among W_L , W_S and sliding window

For the sake of choosing training data, the total user access requests with access log and current user requests are regarded as one user access sequence which is partitioned into data blocks. W_L, W_S and sliding window are respectively set the original value. The original prediction model is constructed with the user access sequence in the sliding window according to certain prediction algorithm. Then the sizes of W_L, W_S and sliding window are adjusted based on prediction precision's changing. The essence of adjusting strategy contains three aspects. First, the sliding window slides ahead and the new user requests are continuous inserted into the sliding window. Second, the prediction model is updated in order to capture the changing user request patterns in time. Third, if the new user access request can compose one new data block, the sizes of W_L, W_S increase one and these new user requests are inserted into two windows. At the same time, some elder data blocks may be deleted from prediction model according to some rules and windows' sizes will change. The concrete adjusting rules are described as following.

1) If the small window's precision is consistent decrease, the sizes of W_L and W_S are shortened and some elder data blocks are deleted from prediction model. The user access requests in the deleted data blocks are obliterated from the sliding window whose length is reduced accordingly. Consistent decrease indicates that any difference between

the prediction precision of the newest data block in the small window and any other is negative. Consistent decrease shows that the outdated access information reduces prediction precision and the prediction model better represents user access characteristics of the elder data blocks which are not consistent with the new user browsing behaviors. Thus the prediction precision of the newest data block falls.

2) If the small window's precision is consistent increase, the width of W_L and W_S is widened. The large window and small window includes the newest data block and their sizes are increased. Consistent increase indicates that any difference between the prediction precision of the newest data block in the small window and any other is plus. Consistent increase denotes the newest user requests enhance prediction capability of the original prediction model so that training data increases.

3) If the small window's precision is stability, the sizes of W_L and W_S are not changed. The precision stability is defined that any difference is very smaller between the prediction precision of the newest data block in the small window and any other. It shows that the newest user requests are consistent with the original model. Thus the large window and the small window cover the newest data block and the eldest data block are deleted from them. At the same time, the corresponding browsing patterns are deleted from the prediction model and the new user requests are added.

4) The prediction model is in a conversion phase if any instance above mentioned does not happen. In order to avoid forgetting the elder training samples too earlier, the wide of the large window is enlarged and the small window's one is kept.

C. Decision algorithm

Suppose that the length of W_L is n and the length of W_S is m , where n is greater than m . The large window's data blocks from the eldest to the newest are respectively labeled from 0 to $n-1$. When a new user request appears, the sliding window goes forward and the new user request is added into it while the bottom of the sliding window does not change. If the new user requests of the sliding window form one new data block n , the changing features of the small window's prediction precision are calculated and the sizes of W_L and W_S are changed according to adjusting rules. Then the length of the sliding window changes and the prediction model's access patterns are updated. In the following section, we specify concrete algorithm and make use of the prediction model which is our previous work [19]. To make this process clear, decision of training data is separated into two steps. First step is the original values of W_L , W_S and the sliding window are respectively set. At the same time, the original prediction model PM is constructed with the user access sequence in the sliding window. Second step is to change training data by adjusting the lengths of different

windows. The following algorithm *DecisionMethod* gives the adjusting strategy.

Algorithm *DecisionMethod*(W_L, W_S, SW, PM, RS)

Input: W_L is the large window, W_S is the small window, SW is the sliding window, PM is prediction model and RS is the new user request sequence.

Output: PM, W_L, W_S, SW

```

BEGIN
  For (every request A of RS)
  BEGIN
    A is inserted into SW and PM
    WHILE (one new data block appears)
    BEGIN
      n=n+1;
      m=m+1; //The sizes of  $W_L$  and  $W_S$  increases one.
      IF (prediction precision of  $W_S$  is consistent decrease)
      BEGIN
        n=n-2;
        //The eldest two data blocks are deleted from  $W_L$ 
        m=n/2; // To change the size of  $W_S$ 
        Every request in the deleted data blocks is deleted
        from SW and PM.
      END
      ELSE
      IF (prediction precision of  $W_S$  is stable)
      BEGIN
        n=n-1; //The eldest data block is deleted from  $W_L$ 
        m=m-1; //To keep the size of two windows
        Every request in the deleted data blocks is deleted
        from SW and PM.
      END
      ELSE
      IF (prediction precision of  $W_S$  is consistent increase)
      BEGIN
        m=n/2;
        ELSE m=m-1; //To increase the large window's size
      END
    END
  END

```

When a new user request appears, we make use of the algorithm in the [20] and its data structure to insert the request into the prediction model so that the changing user behavior patterns are updated in time. When the large windows is shorten, some data blocks are deleted from it and the corresponding user access information is forgotten so that prediction model reduces the outdated browsing patterns and saves space.

IV. EXPERIMENTS

To evaluate our decision method called DM, we adopt Microsoft Visual C++ 6.0 to develop a series of experiments. To compare our method with other, we simulate other system without any training data selection strategy called Non-Selection. DM and Non-Selection both makes use of the prediction model in the [20] during experiments. We compare our approach's performance with Non-Selection from the log day number of training data,

prediction model’s space, prediction precision, hit rate, and traffic incremental rate.

A. Logs and parameters set

We do the trace-driven simulation using two real trace files. One file is from Chinese certain medium-sized education institution’s proxy server log, called CE log. This trace file is collected by one proxy software from January 1, 2005 to January 26, 2005. Every record includes request object’s access information such as IP ,URL and access time. Another file is from American National Lab of Applied Network Research (NLNR) which provides web access logs continuous seven days in one ftp server. We download one proxy server log by authorized username and password, called NLNR_NY, which is composed of continuous user accesses from June 3, 2007 to June 28, 2007.

We remove all dynamically generated files. These files can be in types of “.asp”, “.php”, “.cgi” and so on. We also filter out embedded image files such as “.gif” and “.jpg” because we believe the image file is an embedded file in the HTML file. Access request sequence of each log file is partitioned into user sessions. One user session is one orderly access sequence from the same user. If a user has been idle for more than two hours, we assume that the next request from the same user starts a new user session. We recognize that the time interval of partitioning sessions may introduce some inaccuracy in the simulator, but it will not affect the evaluation of different models.

All of the models make the following configuration. A global model is constructed for all users in each test. All predictions are based on the model. Because of physical systems limitation (e.g. network bandwidth), each model predicts at most a request according to a user’s current request every time. The prefetching cache size is formulated in terms of number of web pages, rather than number of bytes. The approach is more intuitive for interpretation of the results, without altering their significance [16]. The prefetching cache replacement algorithm is LRU. The size of conditional probability threshold affects both hit rate and the amount of traffic increment. A larger threshold allows less data to be prefetched, which is beneficial to traffic, but may decrease hit rate. We take into account a trade-off probability threshold. Thus, conditional probability threshold is set to 0.1.

For our decision method, user access sequence is partitioned into data blocks. Each data block includes one day’s user requests so that CE and NLNR_NY both includes 26 data blocks. Each data block is partitioned into some user sessions according to IP address and time threshold. The large window includes 7 data blocks and the small window’s length is 3 because people regularly browse web every week.

B. Evaluation parameters

We employ the following four metrics [4][21] in the experiments.

Definition 5 Precision is the ratio of the number of correct predictions to the number of total predictions. If users in the subsequent requests access the predicted page that is in the prefetching cache, the prediction is considered to be correct, otherwise it is incorrect. The metric represents the fraction of predicted pages that are actually used.

Definition 6 Hit rate refers to the percentage of user access requests that are found in the prefetching cache.

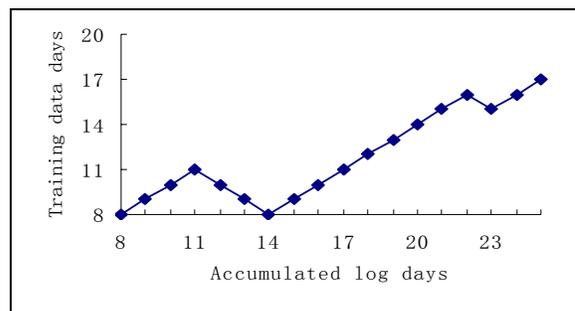
Definition 7 space is the required memory allocation measured by the number of nodes for building a prediction model in the web server for prefetching.

Definition 8 Traffic incremental rate is the ratio of the traffic from undesired pages to the traffic from the total user requests. Some of the prefetched pages will not be actually requested. Therefore, they increase the network traffic overhead.

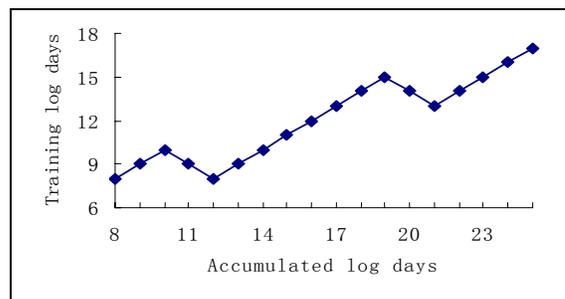
Web prefetching aims at maximizing the first three metrics and minimizing the last one. It is obvious that these metrics are conflicting. The more pages are prefetched, the more probable it is for some of them to be accessed and the hit rate increases. At the same time, precision decreases and network traffic increment is high. Thus, it is a trade-off among these objectives that the model should consider.

C. Decision of training data

In order to decide training data, we respectively choose a part of CE and NLNR_NY to do a series of experiments. Figure 5 depicts the changing process of training data, where abscissa denotes the log’s day number and ordinate is the log’s day number of training data. For example, abscissa is 20, which denotes that 20 days’ log is provided to train prediction model.



(a) Training data versus accumulated log (CE)



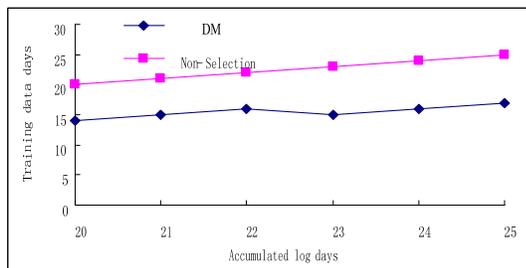
(b) Training data versus accumulated log (NLNR_NY)

Figure 5. Decision of training data

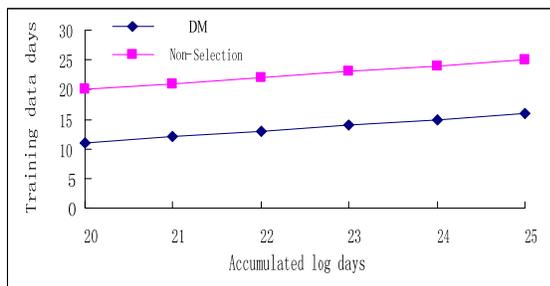
As presented in Figure 5, DM respectively selects the latest 15 days log and 12 days log to construct prediction model for CE and NLANR_NY when the day number of accumulated log is 22. We can conclude that the prediction precision of the small window is consistent increase when the day number of accumulated log varies from 14 to 22 from Figure 5(a) because training data is increasing. At the same time, it also displays that training data decreases when the changing trend of prediction precision is consistent decrease that takes place between days 11 to 14 in Figure 5(a).

D. Comparison of two approaches's training data

In the set of experiments, we display the simulation results of training data comparison among two methods. In the condition of the same day number of accumulated log, Figure 6 gives the log day number of training data with CE and NLANR_NY.



(a) Comparison of training data days using CE log



(a) Comparison of training data days using NLANR_NY log

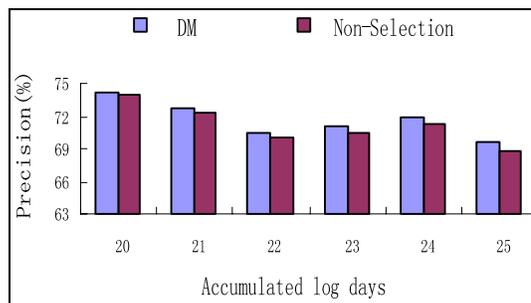
Figure 6. Training data days versus accumulated log days using two logs

As presented in Figure 6, the day number of training data with DM is less than Non-Selection's when the day number of accumulated log varies from 20 to 25. Because the training data in our method is chosen by the changing trend of the small window's precision during constructing prediction and the outdated user requests are deleted from model and the log day number of training data is reduced.

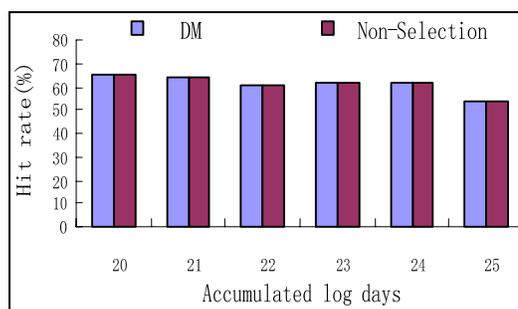
E. Prefetching performance test

We compare the prefetching performance of DM and Non-Selection with two logs in the condition of the same day number of accumulated log. Figure 7 and Figure 8 respectively show different parameter's comparison of prefetching performance using CE and NLANR_NY. In the

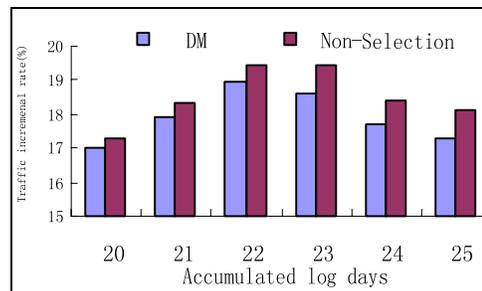
two figures, abscissa is the day number of accumulated log and ordinate respectively represents precision, hit rate and traffic incremental rate for (a), (b), (c) of every figure.



(a) Precision comparison



(b) Hit rate comparison

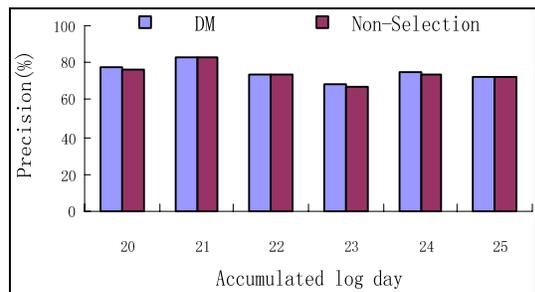


(c) Comparison of traffic incremental rate

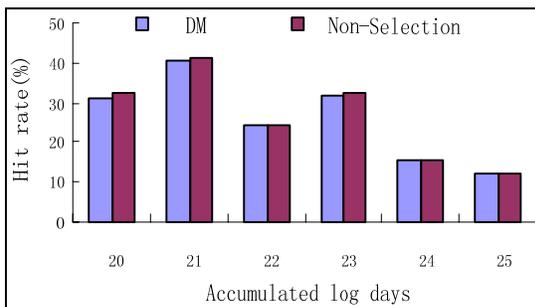
Figure 7. Prefetching performance comparison(CE Log)

In Figure 7, our method and Non-Selection respectively adopt the corresponding result of training data which is displayed in Figure 6 (a). According to the training data, one prediction model is constructed and the succedent one day's user requests are predicted based on the model. For example, DM and Non-Selection respectively use 14 days log and 20 days log to construct the prediction model when abscissa is 20. Then the 21th day's requests are predicted. As presented in Figure 7, the prefetching performance of DM exceeds Non-Selection's. The reason is the Non-Selection method ignores the choosing problem of training data so that the corresponding prediction model does not completely represent the user browsing behaviors. At the same time, DM adopts the technology of adjusting windows to change

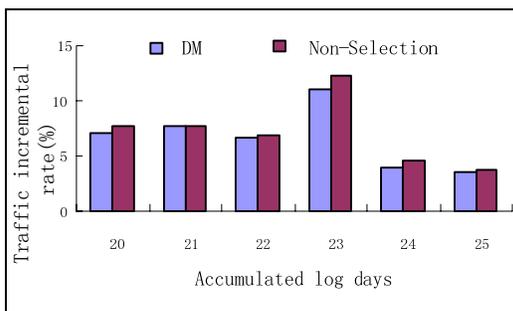
training data so that some outdated access requests are deleted from the prediction model.



(a) Precision comparison



(b) Hit rate comparison



(c) Comparison of traffic incremental rate

Figure 8. Prefetching performance comparison(NLANR_NY Log)

In Figure 8, our method and Non-Selection respectively adopt the corresponding result of training data which is displayed in Figure 6(b). As presented from Figure 8(a) to (c), precision of DM is average higher 0.77% than Non-Selection's and traffic incremental rate is average less 0.49% than Non-Selection's. Although our approach does not have better hit rate or in some cases even worse from the experimental results, it is evident that DM prefetches more web pages correctly than Non-Selection and this is achieved with less cost in the network traffic that has less adverse effect on other network applications. Thus, our algorithm achieves the best performance.

F. Model's space test

We compare prediction model's space of DM with Non-Selection's using two logs in the condition of the same day number of accumulated log. Figure 9 gives the test results

using CE and NLANR_NY, where abscissa represents the day number of accumulated log and ordinate denotes the space reduction rate. It is calculated using the following formula.

$$space\ reduction\ rate = \frac{Non - Selection's\ space - DM's\ space}{Non - Selection's\ space}$$

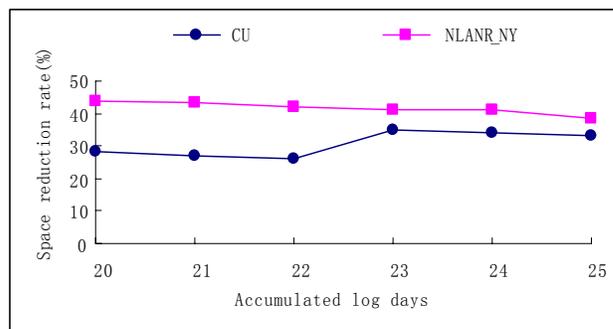


Figure 9. Comparison of prediction model's space using two logs

Figure 9 shows that prediction model's space of DM is always less than Non-Selection's for different logs. The space reduction rate is 25.97% for CE log and it is 41.96% for NLANR_NY log when the day number of accumulated log equals to 22. Figure 9 indicates that our method effectively reduces the prediction model's space.

V. CONCLUSION AND FUTURE WORK

In this paper, we consider the choosing problem of training data and propose a decision method of training data, which is developed according to monitor prediction precision's changing features. It is designed to partition user access sequence into continuous data blocks and makes use of one sliding window, a small window and a large window to capture the precision's characteristics among data blocks so that training data is adjusted. We compare our method with Non-Selection approach from model's space and prefetching performance using two real logs. The experiments show that, for the different day number of accumulated log, our method outperforms Non-Selection's and achieves higher prediction precision with quite low traffic incremental rate and less model's space.

The traces we use are from years ago and some users' behaviors in web surfing could have changed. In the future, we will try to obtain web data from different sources more recent and test the performance of our algorithms.

REFERENCES

[1] J. Domenech, J. A. Gil, et al, "Using Current Web Structure to Improve Prefetching Performance," Science Network, vol. 54, Dec. 2009, pp.1404-1417.
 [2] P. Venketesh, D. R. Venkatesan, and L. Arunprakash, "Semantic Web Prefetching Scheme Using Naive Bayes Classifier," International Journal of Computer Science and Applications, vol. 7, 2010, pp. 66-78.

- [3] B. Ossa, J. Sahuquillo, et al, "An Empirical Study on Maximum Latency Saving in Web prefetching," Proc. IEEE Web Intelligence and Intelligent Agent, IEEE Press, Sep. 2009, pp. 556-559.
- [4] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, "A Data Mining Algorithm for Generalized Web Prefetching," IEEE Transactions on Knowledge and Data Engineering, Sep. 2003, Vol. 15, NO. 5, pp. 1155-1169.
- [5] R. Sarukkai, "Link Prediction and Path Analysis Using Markov Chains," Proceedings of the 9th International World Wide Web Conference, Amsterdam, Holland, May 2000, pp. 377-386.
- [6] L. Shi, Z. Gu, et al, "Popularity-Based Selective Markov Model," Proc. IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, Sep. 2004, pp. 504-507.
- [7] K. Lau and Y. Ng, "A Client-Based Web Prefetching Management System Based on Detection Theory," Lecture Notes in Computer Science, Springer, 2004, vol. 3293, pp. 129-143.
- [8] B. D. Davison, "Learning Web Request Patterns," Web Dynamics: Adapting to Change in Content, Size, Topology and Use. Springer, 2004, pp. 435-460.
- [9] J. Domènech, J. Sahuquillo, et al, "How Current Web Generation Affects Prediction Algorithms Performance," Proceedings of the 13th International Conference on Software, Telecommunications and Computer Networks, Split, Croatia, Sep. 2005.
- [10] Z. Ban, Z. Gu, and Y. Jin, "Selection of Training Period Based on Two-Window," Proceedings of the 10th International Conference on Advanced Communication Technology, IEEE Computer Society Press, Feb. 2008, pp. 2043-204.
- [11] E. Markatos and C. Chronaki, "A Top-10 Approach to Prefetching on the Web," Proceedings of the Eighth Annual Conference of the Internet Society, Geneva, Switzerland, 1998.
- [12] V. N. Padmanabhan and J. C. Mogul, "Using Predictive Prefetching to Improve World Wide Web Latency," Computer Communication Review, 1996, vol. 26, NO. 3, pp. 22-36.
- [13] J. Griffioen and R. Appleton, "Reducing File System Latency Using a Predictive Approach," Proceedings of Summer USENIX Technical Conference, 1994, pp. 197-207.
- [14] J. Borges and M. Levene, "Data Mining of User Navigation Patterns," Proceedings of WEBKDD, 1999, pp. 92-111.
- [15] Z. Su, Q. Yang, et al, "WhatNext: A Prediction System for Web Requests Using N-Gram Sequence Models," Proceedings of the First International Conference on Web Information Systems Engineering, 2000, pp. 200-207.
- [16] T. Palpanas and A. Mendelzon, "Web Prefetching Using Partial Match Prediction" Proceedings of the Fourth Web Caching Workshop, San Diego, California, 1999.
- [17] L. Fan, P. Cao, and Q. Jacobson, "Web Prefetching Between Low-Bandwidth Clients and Proxies: Potential and Performance," Proceedings of the ACM SIGMETRICS'99, Atlanta, Georgia, May 1999.
- [18] M. Deshpande and G. Karypis, "Selective Markov Models for Predicting Web Page Accesses," ACM Transactions on Internet Technology, 2004, vol. 4, NO.2, pp. 163-184.
- [19] C. Bouras, A. Konidaris, and D. Kostoulas, "Predictive Prefetching on the Web and its Potential Impact in the Wide Area," World Wide Web: Internet and Web Information Systems, 2004, vol.7, NO. 2, pp. 143-179.
- [20] Z. Ban, Z. Gu, and Y. Jin, "An Online PPM Prediction Model for Web prefetching," Proceedings of the 9th ACM International Workshop on Web Information and Data Management, Nov. 2007, pp. 89-96.
- [21] J. Domènech, J. A. Gil, et al. "Web Prefetching Performance Metrics: A Survey," Performance Evaluation, 2006, vol. 63, NO. 9, pp. 988-1004.

Towards evolvable Control Modules in an industrial production process

Production control software based on Normalized Systems Theory

Dirk van der Linden¹, Herwig Mannaert², Jan De Laet¹

¹*Electro Mechanics Research Group
Artis University College of Antwerp
Antwerp, Belgium
dirk.vanderlinden, jan.delat@artis.be*

²*Department of Management Information Systems
University of Antwerp
Antwerp, Belgium
herwig.mannaert@ua.ac.be*

Abstract—Normalized Systems theory has recently been proposed to engineer evolvable information systems. This theory includes also a potential of improvement in control software for the automation of production systems. In production control systems, the end user has always the right to have a copy of the source code. However, it is seldom manageable to fluently add changes to these systems, due to the same problems as information systems: couplings, side-effects, combinatorial effects, etc. Finding solutions for these problems include several aspects. Some standards like ISA 88 suggest the use of building blocks on macro level. The OPC UA standard enables these building blocks to communicate and interoperate over the borders of the hosting controllers via local networks or the internet. Consequently, production data collection, manual interfaces and recipe driven production control systems become web service based. Finally the Normalized Systems' theory suggests how these building blocks should be coded on micro level. This paper introduces a control module, based on a design pattern for flexible manufacturing and the principles of Normalized Systems for evolvable software.

Keywords-Normalized Systems, Automation control software, PLC, ISA 88, IEC 61131-3, OPC UA.

I. INTRODUCTION

Industrial communication has in the last 10 years become a key point in modern industry. A continually growing number of manufacturing companies desire, even require, totally integrated systems. This integration extends from electronic automation field devices (PLC: Programmable Logical Controller, PAC: Programmable Automation Controller, DCS: Distributed Control System) to Human Machine Interfaces (HMI) culminating into supervision, trending, and alarm software applications (SCADA: Supervisory Control And Data Acquisition and MES: Manufacturing Execution System). Industrial communication is implemented from field management via process management to Enterprise Resource Planning (ERP) applications (business management).

Just like transaction support software and decision support software systems, production automation systems have also a tendency to evolve to integrated systems. Tracking and tracing production data is not only improving the business, in

some cases it is also required by law (e.g., sectors like food and pharmacy). Because of the scope of totally integrated systems (combination of information systems and production systems) the amount of suitable single vendor systems is low or even non-existing. Large vendor companies may offer total integrated solutions, but mostly these solutions are assembled with products with another history (merged companies or SMEs - Small and Medium Enterprises - bought by larger groups). For the engineer, this situation is very similar to a multi-vendor environment.

Globalisation is bringing opportunities for companies who are focussing their target market on small niches, which make part of a totally integrated system. These products can expand single-vendor systems, or can become part of a multi-vendor system. Moreover, strictly single-vendor systems are rather rare in modern industry. Sometimes they are built from scratch, but once improvements or expansions are needed, products of multiple vendors might bring solutions. Over time, single vendor systems often evolve to multi vendor systems. Minor changes, often optimizations or improvements of the original concept, occur short after taking-in-service. Major changes occur when new economical or technological requirements are introduced over time. As a consequence, software projects should not only satisfy the current requirements, but should also support future requirements [1].

The scope of changes in production control systems, or the impact of changes to related modules in a multi-vendor environment is typically smaller than in ERP systems and large supply chain systems. However, there is a similarity of the problem of evolvability [2]. Since the possibilities of industrial communication increases, we anticipate to encounter similar problems like in business information systems. The more the tendency of vertical integration (field devices up to ERP systems) increases, the more the impact of changes on production level can increase. Since OPC UA (interOperability Productivity and Collaboration - Unified Architecture) enables web based communication between field controllers

and all types of software platforms, over local networks or the internet, the amount of combinatorial effects after a change can rise significantly (change propagation). Based on the systems theoretic concept of stability, a software engineering theory is proposed to engineer evolvable information systems [1]. Although the theory was developed towards business information systems, it has an abstract and generic fundament. Consequently, it should be applicable for production automation control systems too.

This paper introduces a proof of principle on how the software of a production control module can be developed following the principles of Normalized Systems. Some developers could recognize parts of this approach, because it needs to be emphasized that each of the Normalized Systems theorems is not completely new, and even relates to the heuristic knowledge of developers. However, formulating this knowledge as theorems that cause combinatorial effects, supports systematic identification of these combinatorial effects so that systems can be built with minimal combinatorial effects [1]. Normalized Systems allows the handling of a business flow of entities like orders, parts or products. For these process-oriented solutions 5 patterns for evolvable software elements are defined [2]. In this paper however, we focus on the control of a piece of physical equipment in an automated production system. The code of an ISA 88 based control module is not process-oriented but equipment-oriented. The focus of this code is not about how a product has to be made, but about how the equipment has to be controlled. Consequently, we need another type of design patterns. Moreover, we need another type of programming languages because of the nature of industrial controllers. In Section II, we will give an overview of industrial standards on which industrial production control modules can be based. These standards include software modelling and design patterns, communication capabilities, and programming languages. In Section III, an evolvable control module is introduced, including a discussion of change drivers. In Section IV, some changes are implemented. We tested in our lab industrial automation the robustness of the control module against these changes. In Section V we evaluated the proof of principle against the principles of Normalized Systems. During this evaluation, the Design Theorems for Software Stability [2] are used as criteria.

II. INDUSTRIAL STANDARDS

Manufacturing operations can be generally classified into one of three different processes: discrete, continuous, and batch. On October 23, 1995, the SP88 committee released the ANSI/ISA-S88.01-1995 standard [3] to guideline the design, control and operation of batch manufacturing plants. The demand of the users for production systems with a high flexibility and a high potential of making product variants, became important. Process engineers focus on how to handle the material flow to meet the specs of the end-

product. Control system experts focus on how to control equipment. To improve the cooperation of both groups, the SP88 committee had isolated equipment from recipes. This provides the possibility of process engineers to make process changes directly, without the help of a control system expert (reducing the setup-costs). This provides also the ability of producing many product-variants with the same installation (increasing the target market). Expensive equipment can be shared by different production units (enough reducing the production costs). This approach opened the way to what has been called "agile manufacturing". The utilization of ISA 88 data models simplify the design process considerably [7].

Despite the useful ISA 88 terminology and models to structure flexible manufacturing, different interpretations are possible. The standard does not specify how the abstract models should be applied in real applications. Implementers sometimes develop recipes and procedures, which are far more complex than necessary. Since 1995 there have been many applications and a commonly accepted method for implementing the standard has emerged. The S88 design patterns [5] of Dennis Brandl (2007) address this. These patterns might decrease the tension of implementers to make their recipes and procedures more complex than necessary. Unfortunately the part, which describes the connection between computer network systems and control network systems, is limited.

This is where the OPC interfaces come into play. OPC UA is considered one of the most promising incarnations of WS technology for automation. From the very beginning, OPC UA was intended as system interface, aggregating and propagating data through different application domains. Its design, thus, takes into account that the field of application for industrial communication differs from regular IT communication: embedded automation devices such as PLCs, PACs or DCSs provide another environment for web-based communication than standard PCs.

The fundamental components of OPC UA are different transport mechanisms and a unified data modelling [4]. The transport mechanisms tackle platform independent communication with the possibility of optimization with regard to the involved systems. While communication between industrial controllers or embedded systems may require high speed, business management applications may need high data volumes and firewall friendly transport. As a consequence, two data encoding schemes are defined, named OPC UA Binary and OPC UA XML [9].

Data modelling defines the rules and base building blocks necessary to expose an information model with OPC UA. Rather than support data communication, it facilitates the conversion of data to information. The OPC Foundation avoids the introduction of unnecessary new formalisms. Instead, definitions of complex data based on related industrial standards are encouraged. Examples are FDI (Field Device Integration), EDDL (Electronic Device Description

Language), IEC 61131-3 (PLC programming languages) and ISA 88 (batch control). Basically, an OPC UA information model has nodes and references between nodes. Nodes can contain both online data (instances) and meta data (classes). OPC UA clients can browse through the nodes of an OPC UA server via the references, and gather semantic information about the underlying industrial standards. For clients, it is very powerful to program against these complex data types, it brings a potential of code-reuse.

The lowest level of the ISA 88 control hierarchy is the control module. Control modules perform two primary functions: they provide an interface with the physical devices, and they contain basic control algorithms. Control modules encapsulate basic control algorithms and the I/O interface to the actual physical devices. The most common method of programming control modules is any of the IEC 61131-3 programming languages [6]. This standard specifies the syntax and semantics of a unified suite of programming languages for programmable controllers. These consist of two textual languages, IL (Instruction List, has some similarity with assembler) and ST (Structured Text, has some similarity with C or pascal), and two graphical languages, LD (Ladder Diagram, has some similarity with electrical schemes) and FBD (Function Block Diagram, is based on boolean algebra). Industrial programmable controllers are based on divers, often dedicated, operating systems and vendor-dependent programming environments. Besides, earlier days every controller had its own programming language. The release of IEC 61131-3 addresses the problem of too many different programming languages for similar solutions with controllers of different brands. The code of the proof of principle of this paper is written in these languages. However, we emphasize that using the IEC 61131-3 languages is not enough for a 'best practice' implementation. One of the available modelling concepts to analyse the problem and structuring the solution will considerably improve an implementation [11].

III. EVOLVABLE CONTROL MODULES

An invalid function call because the function meanwhile has an updated parameter set is an issue what happens in PLC programming as well as in IT software. Other problems on evolving software occur as well. One of the most annoying problems an automation service engineer confronts is the fear to cause side-effects with an intervention. They have often no clear view on how many places they have to adapt code to be consistent with the consequences of a change. Some development environments provide tools like cross references to address this, but the behaviour of a development environment is vendor-dependent, although the languages are typically based on IEC 61131-3.

In this section we introduce a control module for a motor. We aim to make this motor control software module as generic as possible. In stead of introducing new formalisms,

we based our proof of principle on existing standards. For the modelling, we used concepts of ISA 88 (IEC 61512), for interfacing, we used OPC UA (IEC 62541), and for coding we used IEC 61131-3. More specific, we used the S88 design patterns [5] (derived from ISA 88) because these patterns can be used not only in batch control, but also for discrete and continue manufacturing. None of these standards contain suggestions on how the internal code of a control module should be structured. We introduce a granular structure following the theorems of Normalized Systems. Since the process-oriented approach of evolvable elements for business software [2] is not applicable in our equipment-oriented controller code, we neglect these patterns and use design patterns based on ISA 88 [5]. Every task (action), which must be done by the control module, is coded in a separated POU (Program Organization Unit, sort of subroutine [6]).

In the most elementary form control modules are device drivers, but they provide extra functions like manual/automatic mode, interlocking (permissions), alarming, simulation, etc. [8]. We used the design pattern of Figure 1. This state machine is very simple, when the control systems powers on, the motor comes in the 'off' state. It can be started and stopped via the 'on' and 'off' commands. Hardware failures can cause the motor to go to the 'failed' state, from where a 'reset' command is needed to return to the 'off' state. The concept of this 'failed' state brings us a very important benefit: process safety. Besides, it forms the base for failure notification [10]. This functionality is implemented in a function block we called 'StateAction'. This function block has only one parameter we called 'Device'. The datatype of this parameter is called 'DeviceDataType'. Only a part of this complex datatype is used in the function block 'StateAction'. We called this part 'StateType' (Figure 2). For every other action like controlling the hardware or handling the modes (see further), we have a similar datatype. All action related datatypes are merged into one overall datatype. Exchange of data between the actions can be done via this DeviceDataType (stamp coupling), however without crossing the borders of the control module.

It is obvious that both commands (arrows) and states of Figure 1 are represented as a boolean value in this datatype. One parameter is passing all the necessary data for performing one task: the state action of the control module.

The primary function of our control module is not fulfilled yet: controlling the physical device, in our example the motor, or more general the device hardware. Controlling the hardware is another task in another function block. This function block is receiving the same single parameter 'Device', but it uses another part. The content of the code and the datatype is again very limited (Figure 3). This is one of the key-points of normalized systems: building the application starting from very small modules, performing only one task.

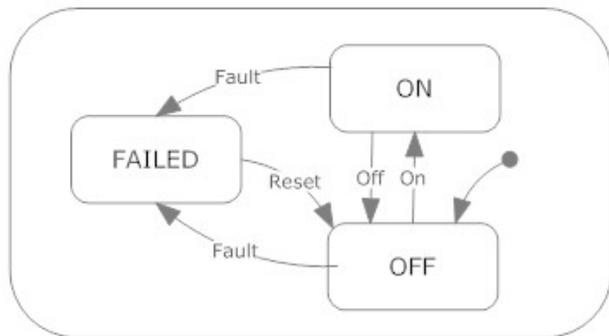


Figure 1: Example of a motor state model [5]

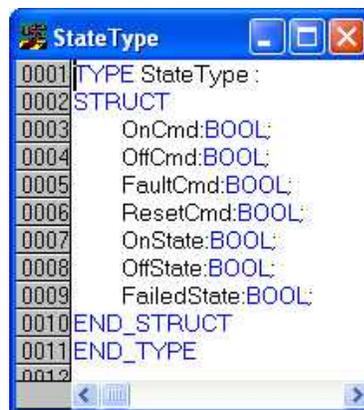


Figure 2: Structure of the datatype 'StateType'

The complex datatype 'DeviceDataType' encapsulates the datatypes 'StateType' and 'HardwareType'. The resulting building block, the Control Module, is connected to only one parameter: an instance of 'DeviceDataType'. This block contains only one action: DeviceAction. This complex action contains two actions: StateAction and HardwareAction, each performing one task. But what is a task? We propose that the definition of tasks can be derived from change drivers. Every interface to our control module can cause or have influence of a change. In our case the change drivers are exposed in Figure 4.

The change driver 'physical equipment' forms the base for the task 'HardwareAction'. The state commands and states forms the base for the task 'StateAction'. Figure 4 suggest another change driver. If we allow low level HMI, the operator becomes the incarnation of a change driver. This means that we add a state machine (ModeAction) to deal with manual/automatic modes, and a subroutine (CommandAction) to separate the commands of the operator (manual commands) and the commands of a higher entity in the automation control project (automatic commands). Following ISA 88 this should be an Equipment Module.

Change drivers have influence on the data structure. To add the functionality above, we need a part 'ModeType' and 'CommandType' in the complex datatype 'DeviceDataType'. So we end up in a Control Module with 4 datatypes (StateType, HardwareType, ModeType, CommandType) encapsulated by the datatype DeviceDataType (Figure 5).

These datatypes are passing all actions, but have all one corresponding action: StateAction, HardwareAction, ModeAction and CommandAction.

IV. ADDING CHANGES

A way to test evolvability is just adding changes and evaluating the impact of these changes. We have 3 actors on our control module: the operator (manual mode), the hardware and the equipment module (automatic mode). Every actor can change his behaviour or can have new expectations or can do new requests. Moreover, one should be able to debug without causing side-effects on other or

older features. In general, we start with a first version. Then we maintain one or more running instances of the control module with initial expected behaviour. Second, we consider the addition of a change, and consequently a possible update of the datatype, existing actions or introduction of a new action. Finally, we make a new instance, check the new functionality and the initial expected behaviour of the older instances as well.

We considered the situation that manual operations could harm automatic procedures. For instance, stopping our motor manually could confuse an algorithm if it is happening during a dosing action. To prevent this, we add the feature manual lock. This means, we still support manual mode, but we disable manual mode during the period a software entity like an equipment module requires this.

Without removing the calls of instances, which dont need this feature, we added a command 'ManLockcmd' to the datatype 'CommandType'. Consequently, this new command becomes part of the overall 'DeviceDataType', so it is passing all actions, but only ModeAction is doing something with this new command.

We considered the situation of a motor instance, which must be able to run in two directions. Again, without removing the calls of instances of single-direction motors, we added a hardware tag 'reverse' to the HardwareType and the commands 'ManReverseCmd' and 'AutoReverseCmd' to the CommandType. In the StateType the tags 'ReverseCmd' and 'ReverseState' were added. As expected, we had to adapt some code in the function blocks 'HardwareAction', 'CommandAction' and 'StateAction'.

On a similar way, we performed other changes like the use of another fieldbus, which required mappings to new hardware addresses. We also introduced a new action SimAction, made (for software testing purposes) to neglect the Fault command (FaultCmd) if no hardware is connected.

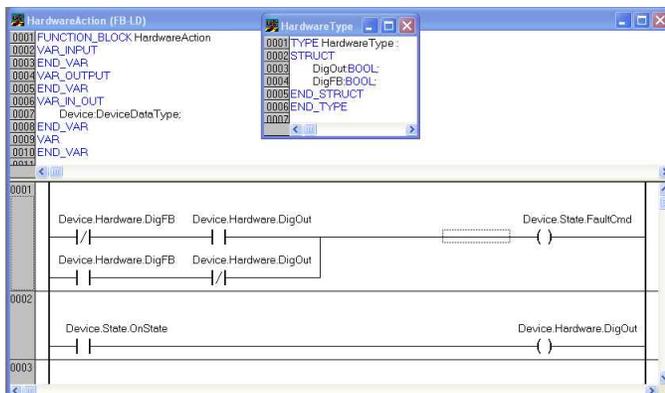


Figure 3: HardwareAction and HardwareType

V. EVALUATION ACCORDING TO THE PRINCIPLES OF NORMALIZED SYSTEMS

Since it is not possible to anticipate on all changes, we cannot test all future change cases. Besides, we are aware that our proof of principle is performed on lab scale, and other, real life situations can occur in a real industrial automation project. In our evaluation we made the assumption that code, following the principles of normalized systems will evolve better than systems, which are not respecting these principles. As a consequence, we checked whether the code is respecting these rules.

First, we consider the separation of concerns. An action entity can only contain a single task. In contrast with industrial usage, where a control module is often just one POU, we made 4 (or more) separate function blocks who are encapsulated by one overall module. For the definition of a task, we based the primary actions on change drivers. Later on, we added the simulation action, which was not directly related to a change driver and thus could be added.

Second, we looked at data version transparency. Data entities that are received as input or produced as output by action entities, need to exhibit version transparency. Only one complex parameter is passed to the control module. Obviously deleting or changing the name of the parameter would destroy existing running connections. Whether adding a parameter would destroy a running instance is vendor-dependent. We stick to one parameter during all changes. Four (later on five) structs are nested. All actions can see the data passing, but every action just picks the data needed for the specific task. We never changed tags, we only added tags. Because of this, earlier instance calls were not affected by data type conflicts.

Third, following the theorem of action version transparency: action entities that are called by other action entities, need to exhibit version transparency. When we changed code, we always beared for not harming the original functionality. For example, we never erased a state or transition in

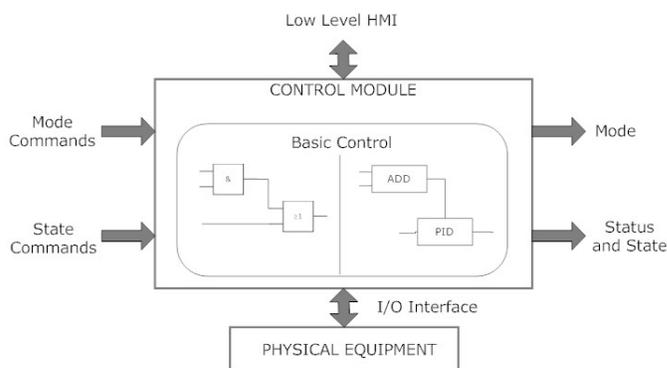


Figure 4: Change drivers of a control module [5]

a state machine, we only added new states and/or transitions. More specific, besides the new functionality, we checked the initial behaviour of existing instances as well.

Fourth, separation of states: the calling of an action entity by another entity needs to exhibit state keeping. It is obvious that the StateAction is managing his state, and keeping it in the related StateType instance. The CommandAction is resulting in a command for the StateAction. The HardwareAction is resulting in a command for the physical hardware. Finally, similar with the StateAction it is obvious that also the ModeAction is managing his state, and keeping it in the related ModeType instance. Besides, the later on added SimAction results in the tags 'SimOnState' and 'SimOffState'.

VI. CONCLUSION AND FUTURE WORK

Evolvability of software systems is important for IT systems, but also a relevant quality value for industrial automation systems. Function blocks of automation systems are programmed close to the processor capabilities. For example, there is a similarity with the IEC 61131-3 language Instruction List (IL) and assembler. The key point of Normalized Systems is a large granularity of software modules, with a structure, which is strictly disciplined to the related theorems. As a consequence, making a proof of principle close to the processor is a very informative exercise to concretize the principles of normalized systems. Besides, this approach can be of great value for improving the quality of industrial automation software projects.

It must be stated that implementing these concepts were highly facilitated by the use of existing industrial standards. They provide us methods to develop the macro-design of software modules, while Normalized Systems provide us guidelines for the micro-design of the actions and data structures encapsulated in these modules. Adding functionality or even adding an action to a (macro) building block, in our case the Control Module, can be done with a limited impact (micro manageable) towards other (macro) entities (bounded impact). To define the most basic actions (tasks)



Figure 5: Structure DeviceDatatype

and data structures, the identification of the change drivers of the concerned entity, in our case represented by the different interfaces to external entities, is essential. This confirms the first theorem for software stability, separation of concerns.

Our future work will be focused on other (macro) elements of ISA 88, which contain other types of control. A Control Module contains mainly basic control, together with a limited coordination control (the mode). We will study on elements with more advanced coordination control code and procedural control, again developed and tested following the principles of Normalized Systems.

Moreover, future work will also be focused on interfaces. Since OPC UA is very generic, we wonder if constraints should be added to the standard to let data communication be compliant to the second theorem of software stability, data version transparency. We wonder whether both currently existing OPC UA transport types, UA binary and UA XML, can be done in a data transparent way.

ACKNOWLEDGMENT

The authors thank Marc Martens (Artesis lab industrial automation), for building the hardware mini-processes needed for performing the testing of this paper, and the good collaboration.

REFERENCES

- [1] van Nuffel Dieter, Mannaert Herwig, de Backer Carlos, Verelst Jan. "Towards a deterministic business process modelling method based on normalized theory" International journal on advances in software - ISSN 1942-2628 -3:1/2(2010), p. 54-69
- [2] Mannaert Herwig, Verelst Jan. "Normalized Systems Re-creating Information Technology Based on Laws for Software Evolvability" Koppa, 2009.
- [3] ANSI/ISA-88.01-1995, Batch Control Part 1: "Models and Terminology."
- [4] Mahnke Wolfgang, Leitner Stefan-Helmut, Damm Matthias. "OPC Unified Architecture", Springer, 2009.
- [5] Brandl Dennis. "Design patterns for flexible manufacturing", ISA, 2007.
- [6] International Electrotechnical Commission (IEC). "IEC 61131-3, Programmable controllers- part 3: Programming languages", Edition 2.0, 2003-01.
- [7] Juoku Virta, Ilkz Seilonen, Antti Tuomi, Kari Koskinen. "SOA-Based Integration for Batch Process Management with OPC UA and ISA-88/95", 15th IEEE International Conference on Emerging Technologies and Factory Automation, september 13-16, 2010, Bilbao, Spain.
- [8] Larry Lamb, Jim Parshall. "Applying S88 - Batch Control from User's Perspective", ISA, 2000
- [9] OPC Foundation. "OPC Unified Architecture, Part1: Overview and Concepts", Release 1.01, february 2009.
- [10] Clark Case, Rockwell Automation. "Applying ISA S88 to Small, Simple Processes", World Batch Forum conference 13-15 nov 2006, Zemst, Belgium
- [11] David Friedrich, Birgit Vogel-heuser. "Benefit of system modeling in automation and control education", American Control Conference, 2007, New York City, USA.

CincoSecurity: A Reusable Security Module Based on Fine Grained Roles and Security Profiles for Java EE Applications

María Consuelo Franky

Department of Systems Engineering
Pontificia Universidad Javeriana
Bogotá, Colombia
lfranky@javeriana.edu.co

Victor Manuel Toro C.

Department of Systems and Computing Engineering
Universidad de los Andes
Bogotá, Colombia
vm.toro815@uniandes.edu.co

Abstract— Almost every software system must include a security module to authenticate users and to authorize which elements of the system can be accessed by each user. This paper describes a reusable security software module that follows the Role Based Access Control model (RBAC), but implementing fine grained roles and grouping them into “security profiles”. This leads to a great flexibility to configure the security of an application by selecting the operations allowed to each profile, and later, by registering the users in one or several of these profiles. The security module has been designed and developed to be the initial code baseline for the development of any Use Cases oriented Java EE system, offering from the beginning a flexible, extensible and administrable access control to the elements of the application.

Keywords-Security; Access control; RBAC; Framework; Java EE; Seam.

I. INTRODUCTION

This paper summarizes the experience of the authors designing and developing a reusable security module, called CincoSecurity, that has been used for several years to control access to the elements of web applications written in Java Enterprise Edition (J2EE initially [8] and later Java EE 5 [9]). Currently, the module CincoSecurity is available [17] under the GPL license, and is used by some software houses in Colombia.

CincoSecurity implements a RBAC (Role-Based Access Control) [6], providing high flexibility to control access to the various elements of a Web application, such as the invocation of an operation of a business component, the access to a web page, and the access to elements within that page. The innovation of CincoSecurity is the use of very fine grained roles, each role having a single permission associated with the invocation of an operation of a business component. From these fine roles—which can be directly controlled by the application server—CincoSecurity allows to define “security profiles” as sets of fine grained roles. This facility of security profiles gives a great flexibility for configuring the security of an application by selecting the operations allowed to each profile, and later, by registering the users in one or several of these profiles.

Any Java EE web application that is to be constructed with the Seam framework [11] gets the following benefits by

integrating the CincoSecurity module: user authentication, registration in the application server of the fine roles derived from the security profiles the user belongs to, and dynamic construction of a personalized menu containing only the entries leading to the use cases allowed for the user. Additionally, CincoSecurity contributes to the application being constructed with use cases to administer the security profiles, to manage user registration in these security profiles and to administer passwords. Additionally, CincoSecurity comes with use cases to register new modules, new use cases and new services, as they become available during a development project, for their security to be administrable.

In the following section, this paper presents the RBAC security model on which the CincoSecurity module is based, and more specifically, the RBAC model applied to the context of Java Application Servers. Then, additional concepts provided by the CincoSecurity module are introduced, as well as its entities model. Later, there is a description of the use cases coming with the CincoSecurity module (e.g., create a new user, create/edit a security profile, add/delete users from a security profile, etc.). At the end, the paper provides a short summary and references to the detailed guidelines [18] for integrating the CincoSecurity module to a Java EE application built with the Seam framework [11]. Finally, there is a comparison with other works, followed by the conclusion and a short description of our future work.

II. EVOLUTION OF THE RBAC SECURITY MODEL

The RBAC model introduced the concept of “role” to control the access to computing resources. The RBAC term was first proposed by Ferraiolo and Kuhn [2], based on previous works of Baldwin [1]. The initial proposal of this model creates a role for each type of job within an organization (cashier, customer service person, office director, ...). Then, each role is assigned with the set of access permissions that are required for this type of job. Finally, each user is enrolled into one or more roles (rather than to specific permissions). This model simplifies security management because the roles (with their associated permissions) tend to be stable, and users can be added or retired easily from roles. The RBAC model allows reinforcing the “least privilege” principle by giving each user

the minimum set of permissions required to perform his work, by enrolling him only in the appropriate roles [6].

From the initial RBAC model (called Core RBAC) the work of Sandhu and colleagues [3] defined extended models, such as the hierarchical RBAC (to include role hierarchy with inheritance of permissions), and constrained RBAC (to prevent, for example, to assign a user to 2 conflicting roles, or to restrict the time slot in which a user can use the permissions of one of its roles).

The main applications of the RBAC model have been in Data Base management Systems, enterprise security management systems, and Web applications that run on application servers [5][6][7].

The wide spread of RBAC models, implemented in numerous products from many providers, led to define an ANSI standard [4] in 2004, aiming to standardize terminology, promote its adoption and improve productivity. However, the current RBAC ANSI standard (consisting of a reference model and a functional specification) has some limitations and gaps as indicated in the work of Bertino and colleagues [7].

III. THE RBAC MODEL APPLIED TO JAVA EE APPLICATION SERVERS

Since the late 90's, the emergence of Application Servers brought a new way to build web applications (both in the enterprise Java platform and in Microsoft .NET), with business components managed by containers that provides added services for security, transaction management, parallelism, pool of connections, logging, etc. [8]

Regarding security, Java EE Application Servers [9] implement the Core RBAC model [6] to control the access to resources based on the roles the user belongs to. In order to take advantage of these security services (and not to write code in the application to internally control access to resources), it is necessary to specify the roles of the application, the association of resources to roles, and the association of users to roles.

A. Enrolling users in roles

In a Java EE application that uses a database to store the authentication and authorization information, the following entities EJB3 (Enterprise Java Beans - version 3) are required [10]:

- An entity "User" shall be implemented (with its corresponding support table in the database), to store users and passwords.
- Entities shall be implemented (with its support tables in the database) to specify the association of each user with one or more roles.
- A "User management" use case shall be implemented to enroll a user in one or more roles.

These facilities are included in CincoSecurity. Similarly, it is also possible to manage users, passwords and roles in a LDAP (Lightweight Directory Access Protocol) server.

B. Controlling access to resources

Seam is a framework to develop Java EE applications, that is being developed by JBoss since 2005, whose principal

author is Gavin King [11][13]. Seam allows to directly expose and use in the Web layer the entities and business components of the application. This simplifies enormously the development by eliminating the intermediaries and conversions between the layers of the application. Seam has been widely accepted and has been incorporated in the recent Java EE 6 standard, under the name of "Web Beans".

To control access to resources in a Java EE application that uses the Seam framework, the following strategies are required [12]:

- An annotation is used to protect each method of the session EJB3. This annotation indicates which roles are authorized to invoke the method.
- The url of each JSF (JavaServer Faces) web page can be protected in the navigation flow descriptor (pages.xml) so that it can be accessed only by users belonging to one of the specified roles.
- Each button or element of a JSF web page can be protected so that it is rendered only to users belonging to one of the specified roles.

C. User authentication

In a Java EE application that uses Seam, the authentication service must be specified in the descriptor components.xml. This service shall be a method of a class of the application, and must implement a query in JPQL (Java Persistence Query Language) [10] to verify the user's password (alternatively this process can also be performed with a LDAP server).

Additionally, the authentication service must also obtain the roles of the authenticated user. With the Seam component called "Identity" these roles can be added to the session and informed to the application server.

D. Controlling access to a JSF page

The application server verifies if the user belongs to a role allowed to access the requested JSF page, and if so, the page is displayed. Similarly, inside the page only the elements that the user is authorized to see are shown (elements such as buttons and text boxes can specify with the attribute "rendered" what roles can see them).

E. Authorizing an action from a JSF page

In a JSF page a button's action is typically associated with the invocation of a method of an session EJB3. The server verifies that the user roles allow him to invoke the associated method (assuming that the method is protected by an annotation indicating the roles that can invoke it).

IV. ADDITIONAL CONCEPTS IMPLEMENTED BY THE CINCOSECURITY MODULE

In addition to the security concepts of Java EE applications that use the Seam framework [11][12], the CincoSecurity module implements additional concepts to provide greater flexibility to define the permissions for user.

A. Use case and services

Definition: A use case is a system's capacity to deliver a useful and indivisible result to the user.

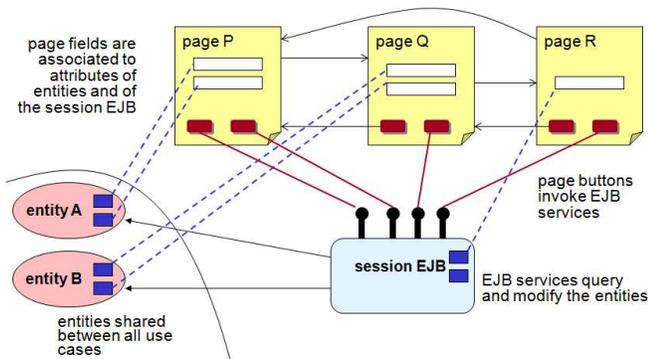


Figure 1: Elements of a Use Case implemented in Java EE with Seam

Definition in terms of its implementation in Java EE with Seam (see Figure 1): a use case consists of one (or more) business entities and a group of services (actions) that act upon them. These actions are implemented as methods of a session EJB3. One or more JSF pages display attributes of the entities involved in the use case, and attributes of the session EJB3 controlling it. In those JSF pages there are actions that invoke the services of the session EJB3. These EJB3 methods are programmed in terms of queries and modifications to the persistent business entities. The navigation flow descriptor contains rules to decide the next page to display.

B. Module

Definition: A module is a set of related use cases. The CincoSecurity module comes with the following use cases, that will be explained below: security profiles management, users management, change of password, basic security reports, registration of menu entries, and registration of modules, use cases and services.

C. Fine grained roles

The CincoSecurity module works with fine grained security roles:

- A role for entering to each use case. The name assigned to this role is the same name of the use case (which is also the Seam name of the session EJB3 that supports the use case).
- A role to invoke each service within a use case (i.e., each of the methods of the session EJB3 that supports the use case). The name assigned to this role is “use case name”_”name of the method”.

D. Protection of resources

- Each session EJB3 that supports a use case is protected with an annotation indicating the role for entering to the use case. For example, the profileGestion use case is supported by the session EJB3 ProfileGestionAction.java (which implements the interface ProfileGestion.java); this EJB3 has the Seam name “profileGestion”. Consequently, the role for entering to this use case is called “**profileGestion**” and the EJB3 class will have the following annotation:
`@Restrict("#{s:hasRole('profileGestion')}")`

- Each service (method) of the session EJB3 is protected by an annotation indicating the role associated with the service. The name of this role is the concatenation of the use case name with the name of the service (with “_” between). For example, the **update** method of the session EJB3 that supports the use case **profileGestion** will have the following annotation:
`@Restrict("#{s:hasRole('profileGestion_update')}")`

- Access to each JSF page of a use case is protected in the navigation flow descriptor by the role for entering to the use case. For example, the page **profiles.xhtml** of the use case **profileGestion** has a navigation flow descriptor called **profiles.page.xml** which contains the following restriction:
`<restrict>#{s:hasRole('profileGestion')} </restrict>`

- Each button in a JSF page should be displayed only to users with the role associated to invoke the action of the button, which corresponds to an EJB3 service (method). For example, the **profiles.xhtml** page of **profileGestion** use case contains a button whose associated action is to invoke the **update** method of the session EJB3 that supports the use case. Consequently the button tag indicates that only it is showed to the role **profileGestion_update**:

```
<h:commandButton id="update"
value="Update" styleClass="button"
action="#{profileGestion.update}"
rendered="#{s:hasRole('profileGestion_update')}"/>
```

E. Security profile

A security profile is a set of fine roles, each fine role expressing the right to invoke a service belonging to a use case. Unlike the role, the concept of security profile is not supported directly by application servers and must be built with additional entities.

The use cases of the CincoSecurity module allow the association of users to roles via security profiles:

- A user can be enrolled in one or more security profiles, so he/she will have the set of fine roles allowed by the union of these profiles.
- There is a *many-to-many* relationship between users and security profiles.
- There is a *many-to-many* relationship between security profiles and roles.

F. Actions after a user authentication

After a user is authenticated, CincoSecurity calculates all the fine grained roles from the security profiles the user belongs to, and informs them to the application server (by assigning these roles to a Seam component called “Identity”). Additionally, the following actions are performed by a EJB3 Login:

- The session timeout is set, according to the parameters stored in the database.
- The user’s menu is built, containing only the entries leading to use cases allowed to the user.
- The security information of the user is added to the session context, should the application logic needs it.

It is important to remark that the access to use cases not authorized to a user by any profile is prevented in two ways. From one side, not authorized use cases do not appear in the user’s menu. From the other side –even if the user types in the url of a not authorized use case– the application server throws a security exception because the fine role for entering to this use case was not included in the list of fine roles that was informed to the application server.



Figure 2: User menu allowing access to all use cases of CincoSecurity

The screen snapshot of Figure 2 shows the menu of a user that is enrolled in security profiles allowing access to all the use cases of the CincoSecurity module.



Figure 3: User menu allowing access to fewer use cases of CincoSecurity

The screen snapshot of Figure 3 shows the menu of another user that is enrolled in security profiles allowing access to just a few use cases of the CincoSecurity module.

V. ENTITIES MODEL OF THE CINCOSECURITY MODULE

The entities model shown in Figure 4 illustrate the relationship one-to-many from Module to Usecase, and from Usecase to Service.

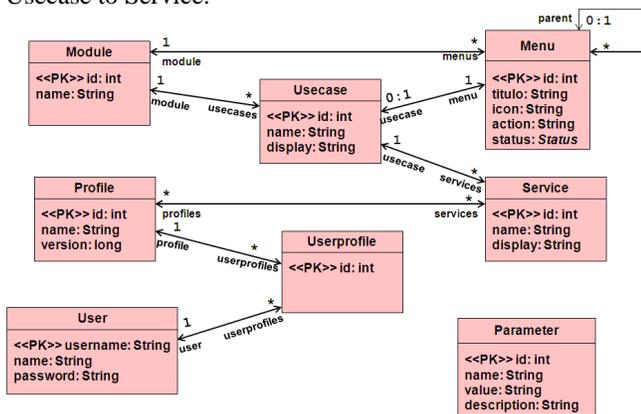


Figure 4: Model of entities of the CincoSecurity module

Figure 4 also illustrates the relationship many-to-many between Profile (security profile) and Service, as well as between Profile and User (via the intermediate entity Userprofile). Each menu entry may have submenus (only terminal menu entries have an action for going to the entry page of a use case).

The system parameters are arbitrary. They can be used, for example, to record the session timeout, the path of the directory to store reports, the address of printers, etc.

VI. USE CASES OFFERED BY THE CINCOSECURITY MODULE

The following are the use cases offered by the CincoSecurity module:

A. CRUD Use cases

The CincoSecurity module offers:

- A use case to list/add/edit and remove **parameters** of the application.
- A use case to list/add/edit and remove **modules** of the application.
- A use case to list/add/edit and remove the **use cases** of a module.
- A use case to list/add/edit and remove the **services** of an application’s use case.
- A use case to list/add/edit and remove **menu entries**.

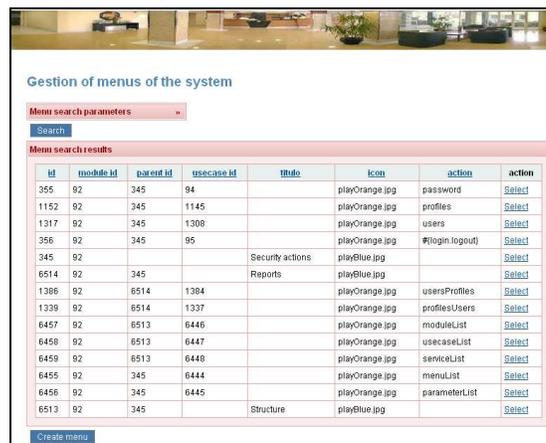


Figure 5: Use case to list/add/edit or remove menu entries.

It can be easily specified which menu entries have a submenu, as well as the use case associated with a terminal entry (see Figure 5).

B. Management of security profiles

This use case allows to add/edit/remove security profiles. Initially, the existing security profiles are listed. When a security profile is selected, the modules, use cases and allowed services are shown, so that the user can check or uncheck services (see Figure 6 in the next page).

Similarly, the user can create a new security profile. In this case, the system displays all modules, and within it, the use cases and services, for the user to select those allowed by the new profile.

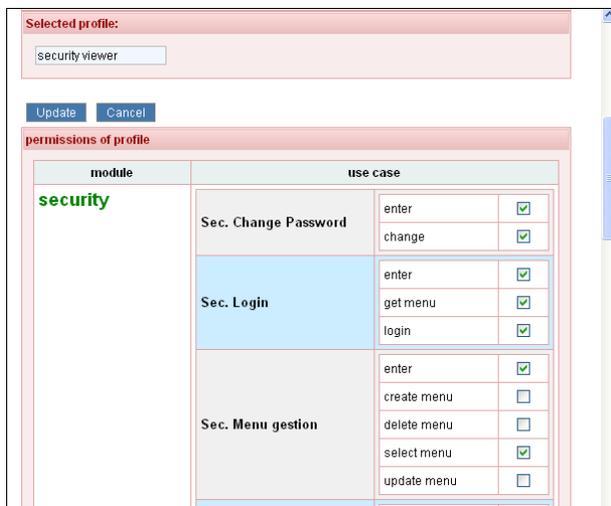


Figure 6: Use case to manage security profiles

C. Management of users

This use case allows to add users of the application, indicating its name, login and password. It also allows to enroll the new user in one or more security profiles.

D. Password change

This use case allows a user to change his password. Passwords are stored encrypted.

E. Report of security profiles vs users

This use case reports, for each security profile, which users are enrolled.

F. Report of users vs security profiles

This use case reports, for each user, in which security profiles is enrolled.

VII. HOW TO INTEGRATE THE CINCOSECURITY MODULE TO A JAVA EE SEAM APPLICATION

The CincoSecurity module is open source with GPL License [16]. It can be downloaded from SourceForge [17] in the form of an Eclipse project [15]. It comes ready to be deployed on the application server JBoss [14], but can be installed in any other Java EE application server by following the guidelines of the Seam manual [13].

The documentation accompanying the CincoSecurity module explains in detail how to deploy and execute the module, and how to integrate it with a Java EE application built with Seam. In particular, detailed explanations are included for registering application's use cases and services in the security module. An earlier publication about the CincoSecurity module, oriented to programmers, focused on these technical details [18].

It is important to emphasize that to incorporate and manage the security of an application, the modules of the application, the use cases contained in such modules, and the services offered by these use cases must be registered into the CincoSecurity module (by using CincoSecurity's use cases provided for this). This way, the fine roles associated

with these services can be included in the security profiles, and the access to these use cases will appear in the menu of authorized users.

VIII. COMPARISON WITH OTHER WORKS

There are other security modules proposed for the Java EE technology. Currently, in the SourceForge portal there is a dozen of software projects related with security for Java EE 5 [9] applications, but most of them are proposals without implementation (i.e., they are in planning status). Two relevant projects with implementation and good acceptance from users are the following:

- *JPA Security* [19] is an Access Control Solution for the Java Persistence API (JPA) [10] with support for role-based access control, access control lists (ACLs) and domain-driven access control. Compared with CincoSecurity, this project does not offer access control to web pages as CincoSecurity does. JPA Security uses access rules in terms of database operations, which provides a different type of flexibility from CincoSecurity, where security profiles are defined in terms of the services offered by the use cases of the application.
- *[fleXive]* [20] is a Java EE 5 open-source framework for the development of complex and evolving web applications. It offers an administration module that manages users and security. It implements an access control list based approach, combined with roles. Compared with CincoSecurity, this project uses 10 coarse roles with predefined permissions related to the administration module, while CincoSecurity lets to define any number of security profiles, each one as a set of fine roles related to the services of the application (not only to the security services). We believe it is more intuitive to associate the users to these security profiles and not to the [fleXive] Access control lists (ACL), that define lists of permissions attached to arbitrary objects. [fleXive] does not offer access control to elements of web pages, as CincoSecurity does.

With respect to recent proposals for extending the RBAC model, some research works as [21][22] try to statically validate the correctness of roles usage in an application, for solving what they call the fragility of traditional dynamic checks. CincoSecurity does not implement static checks, but its strategy of fine grained roles enables to automate the correct incorporation of security in a web application. In effect, by following the names discipline explained in this paper it is possible to automatically add annotations to each method of the session EJB3 controlling a use case, in order to permit its access only to users having the associated fine role; it is also possible to automatically modify button tags of JSF pages for rendering it only to users having the corresponding fine role.

On the other hand, it is important to note that Seam offers a complete security module [13], that is based on (coarse) roles, permissions and rules, that achieves a very flexible control of resources. The CincoSecurity module takes advantage of the Seam security by using some of its facilities

related with authentication, restriction annotations for roles, and tags of JSF pages for rendering only for the appropriate role. However, we believe that for an administrator it is more difficult to write rules for granting fine permissions to roles (as is done in the Seam module), than to configure security profiles by checking the services of the application to be granted (as is done in the CincoSecurity module). Also, given that CincoSecurity does not use permissions nor rules (only fine grained roles), the incorporation of security to a web application can be automated, as it was explained above; with the Seam module it seems more difficult to automate the incorporation of security.

IX. CONCLUSION AND FUTURE WORK

Needless to say that developing a secure Java EE application is a difficult job, where dozens of subtle details must be handled coherently. The CincoSecurity module provides a complete code baseline to develop a Java EE application with the Seam framework, incorporating from the beginning full and flexible access control to the use cases and services of the application being developed. The CincoSecurity module also provides the use cases required to administer the users and their access permissions to the use cases and services of the application being developed.

With respect to the Core RBAC model [6], an access permission is materialized in CincoSecurity as the right to invoke a method (service) of a business component. Fine grained roles are defined and implemented, each one having just one permission to invoke a single method (service) of a business component (session EJB3). There is also a fine grained role to allow entrance to each application's use case, as well as a fine grained role to grant access to each one of the services provided by the use case. The concept of "security profile" is defined and implemented as a set of fine grained roles.

The CincoSecurity module takes advantage of the low level access control principle implemented by any application server, by feeding the application server with the fine grained roles included in the security profiles the authenticated user belongs to. Additionally, the CincoSecurity module dynamically builds a customized menu containing only the entries leading to the application's use cases authorized for the user.

The CincoSecurity module is a Java EE 5 application built using the Seam framework [11][12]. It is distributed under the GPL license and can be freely downloaded from <http://sourceforge.net/projects/cincosecurity>. CincoSecurity is used by several software houses in Colombia.

As future work, CincoSecurity will be extended to automate the incorporation of security to a web application (business components and web pages), as well as to include other capabilities of the Seam security module, like Identity management, in a compatible way with our approach.

REFERENCES

- [1] R. L. Baldwin, "Naming and Grouping Privileges to Simplify Security Management in Large Databases", Proceedings of the 1990 IEEE Symposium on Research in Security and Privacy (Oakland, CA), IEEE Computer Society Press, pp. 116-132, 1990.
- [2] D. F. Ferraiolo and D. R. Kuhn, "Role-Based Access Control". Proc. 15th Nat'l Information Systems Security Conf., Diane Publishing Company, pp. 554-563, 1992.
- [3] R. Sandhu, C. L. Feinstein, and C. E. Youman, "Role-Based Access Control Models". IEEE Computer Magazine, pp. 38-47, 1996.
- [4] American National Standard for Information Technology – Role Based Access Control, ANSI INCITS 359-2004, 2004.
- [5] B. Messaoud, "Access Control Systems: Security, Identity Management and Trust Models", Springer Science+Business Media, Inc., 2006.
- [6] D. F. Ferraiolo, D. R. Kuhn, and R. Chandramouli, "Role Based Access Control", Artech House 2003, 2nd Edition 2007.
- [7] N. Li, J. W. Byun, and E. Bertino, "A Critique of the ANSI Standard on Role-Based Access Control", IEEE Security and Privacy, Volume 5, Issue 6, pp. 41-49, 2007.
- [8] D. Alur, J. Crupi, and D. Malks, "Core J2EE Patterns: best practices and Design Strategies", Sun Microsystems - Prentice Hall, 2001.
- [9] Sun Microsystems, "The Java EE 5 Tutorial", http://java.sun.com/javaee/5/docs/tutorial/doc_01.18.2011
- [10] M. Keith and M. Schincariol, "Pro EJB 3: Java Persistence API", Apress, 2006.
- [11] M. Yuan and T. Heute, "JBoss Seam: Simplicity and Power Beyond Java EE", Prentice Hall, 2007.
- [12] D. Allen, "Seam in Action", Manning Publications Co., 2009.
- [13] JBoss Seam Group, "Reference manuals of JBoss Seam", http://seamframework.org_01.18.2011
- [14] JBoss Community, "JBoss Application Server", http://www.jboss.org/jbossas_01.18.2011
- [15] Eclipse Open source development platform comprised of extensible frameworks, tools and runtimes, http://www.eclipse.org_01.18.2011
- [16] General Public License, http://www.gnu.org/licenses/gpl.html_01.18.2011
- [17] M. C. Franky, V. M. Toro, and R. López, "CincoSecurity Module", http://sourceforge.net/projects/cincosecurity_01.18.2011
- [18] M. C. Franky, V. M. Toro, and R. López, "CincoModule: Módulo de seguridad basado en roles finos y en perfiles de seguridad para aplicaciones Java EE 5". Quinto Congreso Colombiano de Computación (5CCC), Cartagena-Colombia, Abril 2010. ISBN: 978-958-8387-40-6.
- [19] "JPA Security", http://jpasecurity.sourceforge.net_01.18.2011
- [20] D. Lichtenberger, M. Plessner, G. Glos, J. Wernig-Pichler, H. Bacher, A. Zrzavy, and C. Blasnik, "[fleXive]TM 3.1 Reference Documentation", Copyright © 1999-2010 UCS - unique computing solutions gmbh, http://www.flexive.org/docs/3.1/xhtml/index.xhtml_01.18.2011
- [21] P. Centonze, G. Naumovich, S. J. Fink, and Marco Pistoia, "Role-Based access control consistency validation". In Proceedings of the 2006 international symposium on Software testing and analysis (ISSTA'06). ACM, New York, NY, USA, pp. 121-132, 2006
- [22] J. Fischer, D. Marino, R. Majumdar, and T. Millstein, "Fine-Grained Access Control with Object-Sensitive Roles", ECOOP 2009 – OBJECT-ORIENTED PROGRAMMING, Lecture Notes in Computer Science, Volume 5653/2009, pp. 173-194, 2009.

A Personalized Recommender System Model Using Colour-impression-based Image Retrieval and Ranking Method

Ana Šaša, Marjan Krisper
 Faculty of Computer and
 Information Science
 University of Ljubljana
 Ljubljana, Slovenia
 {ana.sasa,
 marjan.krisper}@fri.uni-lj.si

Yasushi Kiyoki,
 Shuichi Kurabayashi,
 Faculty of Environment and
 Information Studies,
 Keio University
 Fujisawa, Kanagawa Japan
 {kiyoki,
 kurabaya}@sfc.keio.ac.jp

Xing Chen
 Department of Information &
 Computer Sciences
 Kanagawa Institute of
 Technology
 Atsugi, Kanagawa
 chen@ic.kanagawa-it.ac.jp

Abstract— This paper points out that achievements in the field of multimedia analysis and retrieval represent an important opportunity for improvement of recommender system mechanisms. Online shopping systems use various recommender systems; however a study of different approaches has shown that they do not exploit the potential of information carried by multimedia product data for product recommendations. We demonstrate how this can be accomplished by a personalized recommender system model that is based on analysis of colour features of product images. We present an approach for extraction of colour-properties of images in order to represent impressions related to the human perception of images. Colour-properties are based on image colour histograms, psychological properties of colours and a learning mechanism. Based on the extracted colour-properties, the method retrieves and ranks the images corresponding to the desired impressions. The architectural framework of the model is based on service-oriented architecture in order to promote its flexibility and reuse, which is important when applying the model to existing recommender system environments. An experimental study was performed for decorative photography domain.

Keywords- *product recommendation; image retrieval; E-business; service-oriented architecture.*

I. INTRODUCTION

Online shops provide different Web-based services for customer-product matching and product data representation in order to support customers at their decisions when buying the products. Customer-product matching mechanisms differ greatly based on the amount of effort the customer has to invest in order to find the desired product. The most basic approach that requires the most effort from the customer is catalogue browsing. As customers are often unable to evaluate all available alternatives in great depth, they tend to use two-stage processes to reach their purchase decisions using product catalogues: firstly they screen a large set of available products and identify a subset of the most promising alternatives; secondly they evaluate the latter in more depth, perform relative comparisons across products on important attributes, and make a purchase decision [1]. Other

types of approaches are known as recommender systems. Recommender systems help customers find the products they would like to purchase by producing a list of recommended products for each given customer [2]. Schafer et al. have developed a taxonomy of e-commerce recommender systems [3][4] based on two key dimensions: a) the degree of automation, which depends on the effort the customer has to invest in order to get the recommendation, and b) the degree of persistence in recommendations, which depends on whether the recommendations are based on data regarding previous customer sessions with the system or not. One of the most common partially automated recommender systems are product search services where customers enter desired product attributes (search query) and search results comprise those products that correspond to the search query or are the closest match for the search query. An example of a completely automated recommender system is an implicit customer-product matching system. They do not require explicit customer actions, and autonomously recommend selected products to individual customers, for example in the side frame of the Web page.

There are many research contributions in the field of recommendation techniques; examples are [2][5][6][7][8][9][10][11]. They range from very basic, such as presentation of the most popular products, to more advanced, such as web mining techniques [5], collaborative filtering [12], decision tree induction [2], association rule mining [13], etc. However, a study of different recommendation techniques ([2],[5-11],[14-21]) that existing recommender systems techniques do not take into consideration information contained in the multimedia product data. The multimedia product data is used mainly to present the product to the customer after the products to be presented have been selected. We observe that there is an important potential to improve the customer-product matching mechanisms using not only text-based data about the products, but also product multimedia data. In the recent years, advances in the field of multimedia have provided several important results and the potential is there to improve existing customer-product matching mechanisms.

In this paper we present an innovative model for improvement of recommender systems, using product image data. We have applied the colour-impression-based image retrieval and ranking method [25] to the design of this model. In [25], the colour-impression-based image retrieval method and its system architecture have been proposed for realizing image-colour processing dealing with emotion-aspects of human senses to images. The purpose of our model is not to replace the existing recommender system techniques, but to enhance them by utilizing image based information about products. The presented model uses a method for product image analysis which can automatically determine colour-properties of product images. The extracted colour-properties are used in order to improve the product search results and product recommendations. Colour-properties are extracted based on image colour histograms, psychological properties of colours and a learning mechanism. The method that implements this behaviour has been presented as the colour-impression-based image retrieval and ranking method in [24] and [25]. As far as we know, there is no other product recommendation system that exploits colour-impressions contained within product images in order to improve recommendation results. An important characteristic of this approach is that colour-impressions based on image perception are subjective and pertain to individual customers. Therefore, one of the requirements of this approach is to take into consideration different perceptions of different individuals. The proposed model can be valuable for different companies, especially for those where colour and subjective properties of their products play an important role in product selection for the customer, for example fashion and art domains. We show an example implementation of the model for decorative photography domain.

The remainder of the paper is structured as follows. In section two, we briefly review the colour-impression-based image retrieval and ranking method [24] [25]. In section three, we discuss the image-based recommender system model and how the colour-impression-based image retrieval and ranking method is applied to a recommender system. In section four, we represent an experimental study from the domain of decorative photography. In section five, we give concluding remarks.

II. COLOUR-IMPRESSION-BASED IMAGE RETRIEVAL AND RANKING METHOD [24][25]

In this section, as an image retrieval method for realizing our personalized recommender system model, we briefly review the colour-impression-based image retrieval and ranking method which has been proposed in [24][25]. In this paper, we apply this method to the design of our recommendation system model with product image data. The main feature of the colour-impression-based image retrieval and ranking method is how to deal with the colour features in products. A knowledge base is used to compute the relevance between a specific colour and a set of impression concepts, such as “sharp” and “cool”, by using a dictionary that defines abstract semantics of colours [24][25]. A colour-impression association knowledge base provides a matrix (colour impression definition matrix; CID matrix) that

defines colour features of colour schemas related to 130 colour variations (Figure 1). Each colour schema corresponds to a specific impression in human’s perception [26]. In [25], the colour-impression space has been created, using 120 chromatic colours and 10 monochrome colours defined in the “Colour Image Scale” [27], which is based on the Munsell colour system [28], as shown in Figure 2. Figure 1 shows several examples of the colour-emotions defined in “Colour Image Scale” [27]. The colour-emotions are expressed in 120 chromatic colours and 10 monochrome colours which are based on the Munsell colour system. Each colour schema, which corresponds to a specific emotional perception of humans, is described by using a combination of several colours from the 130 basic colour set. Each colour-emotion, such as vivid(cs2) and sweet(cs5), defines colour features in 182 sets of colour schema related to 130 colour variations. Each colour schema corresponds to a specific emotion in human’s perception.



Figure 1. Colour-impression association for defining colour features in 182 sets of colour schema related to 130 colour variations defined in “Colour Image Scale” [27]. Each colour-impression definition, such as “vigorous (cs1)” defines a set of weight for colours

R/V	YR/V	Y/V	GY/V	G/V	BG/V	B/V	PB/V	P/V	RP/V	N/10
R/S	YR/S	Y/S	GY/S	G/S	BG/S	B/S	PB/S	P/S	RP/S	N/9
R/B	YR/B	Y/B	GY/B	G/B	BG/B	B/B	PB/B	P/B	RP/B	N/8
R/P	YR/P	Y/P	GY/P	G/P	BG/P	B/P	PB/P	P/P	RP/P	N/7
R/Vp	YR/Vp	Y/Vp	GY/Vp	G/Vp	BG/Vp	B/Vp	PB/Vp	P/Vp	RP/Vp	N/6
R/Lgr	YR/Lgr	Y/Lgr	GY/Lgr	G/Lgr	BG/Lgr	B/Lgr	PB/Lgr	P/Lgr	RP/Lgr	N/5
R/L	YR/L	Y/L	GY/L	G/L	BG/L	B/L	PB/L	P/L	RP/L	N/4
R/Gr	YR/Gr	Y/Gr	GY/Gr	G/Gr	BG/Gr	B/Gr	PB/Gr	P/Gr	RP/Gr	N/3
R/Di	YR/Di	Y/Di	GY/Di	G/Di	BG/Di	B/Di	PB/Di	P/Di	RP/Di	N/2
R/Dp	YR/Dp	Y/Dp	GY/Dp	G/Dp	BG/Dp	B/Dp	PB/Dp	P/Dp	RP/Dp	N/1
R/Dk	YR/Dk	Y/Dk	GY/Dk	G/Dk	BG/Dk	B/Dk	PB/Dk	P/Dk	RP/Dk	
R/Dgr	YR/Dgr	Y/Dgr	GY/Dgr	G/Dgr	BG/Dgr	B/Dgr	PB/Dgr	P/Dgr	RP/Dgr	

Figure 2. 130 Munsell Basic Colour Variations defined in “Colour Image Scale” [27].

The system converts RGB colour values to HSV colour values per pixel of each image. HSV is a widely adopted space in image and video retrieval because it describes perceptual and scalable colour relationships. The system clusters HSV pixels into the closest colour of the predefined 130 Munsell basic colours [1], and calculates the percentage of each colour to all pixels of the image. And then the system creates 130 HSV colour histogram. The image metadata generation function $f_{image_metadata}(D)$ is defined as follows:

$$f_{image_metadata}(D) \rightarrow I := \{i_{[0]}, \dots, i_{[129]}\}$$

where D denotes an image data, and $i_{[n]}$ denotes a i -th colour feature value. The method analyzes images to generate a colour-impression based metadata vector. The method extracts the colour-impression features of the image and

inserts them into the metadata cache database. The system analyzes the colour-impression features by performing the following four steps as described in Figure 3:

- **(Step-1)** The system decodes an input image file such as JPEG and PNG.
- **(Step-2)** The system converts RGB colour values to HSV colour values per pixel of each image. HSV is a widely adopted space in image and video retrieval because it describes perceptual and scalable colour relationships.
- **(Step-3)** The system clusters HSV pixels into the closest colour of the predefined 130 Munsell basic colours[1], and calculates the percentage of each colour to all pixels of the image. And then the system creates 130 HSV colour histogram.
- **(Step-4)** The system extracts the colour-impression for the image by correlation calculations between 182 colour schemas (182 impression word sets) and 130 basic HSV colours.

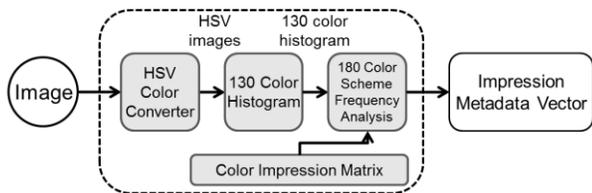


Figure 3. The process of colour-impression based metadata extraction [24][25].

III. ARCHITECTURAL FRAMEWORK OF THE PERSONALIZED RECOMMENDER SYSTEM MODEL USING COLOUR-IMPRESSION-BASED IMAGE RETRIEVAL AND RANKING METHOD

Figure 4 illustrates the architectural framework overview of our personalized recommender system model. It is based on service-oriented architecture (SOA) due to several important SOA characteristics, especially flexibility and reuse. These allow for easier adaptations of the model to different environments. To address the service orchestration principles of SOA [34][35][36][37], the key business process of the online shopping domain (i.e., the Online purchase process), is implemented with the Web Service Business Process Execution Language (BPEL) [38].

A high level overview of the process and its main subprocesses is given in Figure 4 using the standard Business Process Model and Notation 2.0 (BPMN) [39]. The recommender model concerns its Product search and selection subprocess, as shown in Figure 5. It is important to note that the model is generic and that the proposed SOA framework allows that BPEL processes are extended with existing recommender system services. Figure 6 demonstrates the BPMN model of the Product search and selection subprocess and how the architectural components relate with the colour-impression-based image retrieval and ranking method.

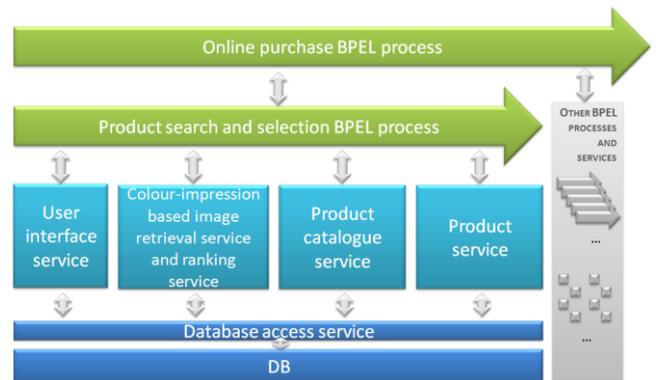


Figure 4. Architectural framework overview

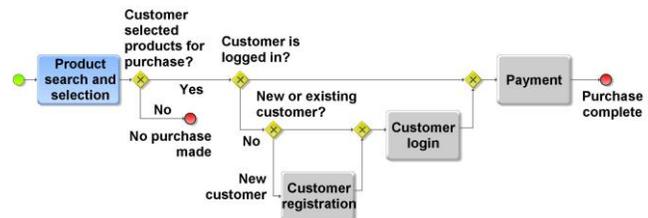


Figure 5. Online purchase process overview

Swimlanes are used to show which services are invoked by the BPEL process for execution of the corresponding process tasks and which tasks are performed by the user. For clarity and understanding of our work, the process models do not show details that are irrelevant for the research presented in this paper. One can observe that the model is based on interactivity with the user and that the presentation of the products to the user is updated for every main user action (*interactivity loop*). There are three main types of user activities when searching for desired products:

- The customer may select impression words appertaining to the desired products. In this case the colour-impression-based image retrieval and ranking function returns products that are the most relevant for the given impression words based on the product images. The presentation of the products for the customer changes and shows the resulting products.
- The customer may select a product, for example to see more details about it. In this case, the colour-impression-based image retrieval and ranking is performed based on the images of the selected product. Products with similar-impression images are returned and are presented to the customer in the side frame.
- The model also allows for the standard product attribute selection, such as type of the product, colour, size, material etc. In case the customer changes the attributes and if the customer has already chosen the impression words, products are first filtered in order to retrieve the products that correspond to the selected attributes. Then the colour-impression-based image retrieval and ranking function performs the search based on the filtered product images. The presentation of the products for the customer changes and the resulting products are shown.

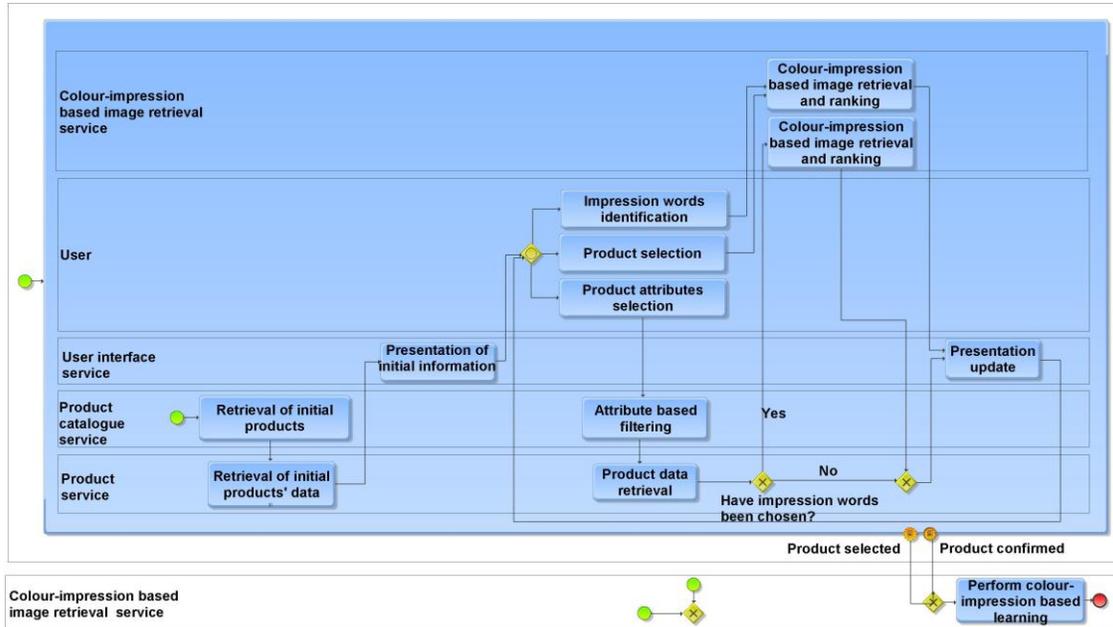


Figure 6. Product search and selection process overview

The colour-impression-based image retrieval and ranking function is integrated in all interactivity loops. However, the process differs if the customer is logged or not. If the customer is logged in, their personal CID matrix is used. Otherwise, the common domain CID matrix is used as the basis.

Colour-impression-based learning is performed every time a user selects the product and every time the user confirms a product for purchase. Based on the impression words specified by the customer and the selected/confirmed products, product image analysis is performed to obtain the colour features of the images. The corresponding CID matrix is adjusted in order to give a stronger association between the impression words and selected product images. If the customer is not logged in and the common domain CID matrix is used, the adjustments are performed on a copy of the common domain CID matrix. If the customer later logs in, the adjustment is applied to the adjusted CID matrix. Thus the CID matrix implements the personalization based for specific customers. Otherwise, the matrix is used only during the customer session.

IV. EXPERIMENTAL STUDY

We have performed an experimental study of the model for the decorative photography domain. The corresponding online purchase business process was implemented based on the architectural framework discussed in the previous section. For the use cases examples presented in the remainder of this section, the selected product attributes were: category: scenery, landscapes, places; size: 4:3 and 16:9. 205 images corresponded to this attribute selection. Table I demonstrates photograph images of three result sets of the colour-impression-based image retrieval and ranking: first for the impression word “earnest”, second for the word “exact”, and third for the words “earnest, exact”. The

impression word query “earnest, exact” returns those images that are semantically the closest to both impression words. Correlation values are given for the resulting product images.

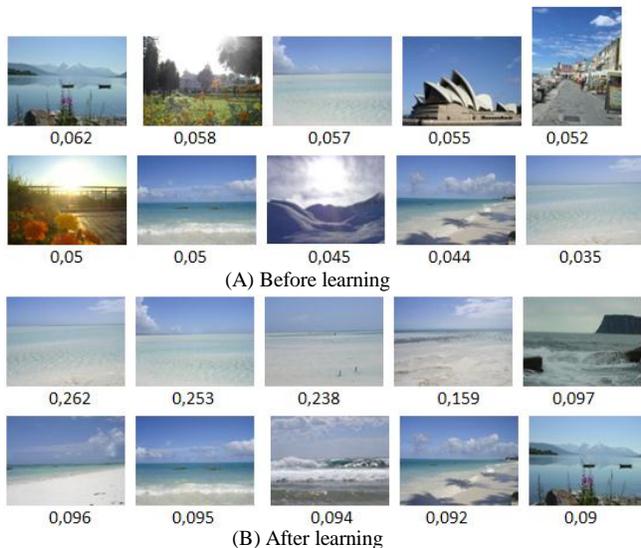
TABLE I. PRODUCT IMAGES PRESENTED TO THE CUSTOMER BASED ON AN AJUSTED CID MATRIX AND THREE EXAMPLE IMPRESSION WORD SETS

Impression words	Five resulting images with the highest correlation values			
Earnest Corresponding colour histogram:	 0,113	 0,081	 0,070	 0,066
Exact Corresponding colour histogram:	 0,085	 0,083	 0,077	 0,076
Earnest, exact	 0,129	 0,129	 0,121	 0,118

Table II demonstrates two image result sets for the impression word “crystalline”. Table II (A) represents the results based on the common domain CID matrix. Let us consider the example where after the customer is presented with the resulting 10 photographs. The customer then selects the product with the last image in the result set (image with the correlation 0,035 in Table II (A)). When the product is selected, colour-impression-based learning is performed. Besides the product details for this photograph, the customer is presented the image result set of the Table II (B) without the first photograph (as it is the same as the selected photograph). One can easily notice the difference in the results and the adaptation to the customer’s perceptions. Furthermore, next time the customer performs the search for the word “crystalline”, the resulting photographs would be

refined and adjusted to the customer. For the same input set of images the result would be as presented by Table II (B).

TABLE II. EXAMPLE IMAGE RESULT SETS FOR THE IMPRESSION WORD "CRYSTALLINE"



V. CONCLUSION AND FUTURE WORK

This paper has presented a recommender system model based on the colour-impression-based image retrieval and ranking method in [24][25]. This paper has demonstrated how the colour-impression-based image retrieval and ranking method can be applied to recommender systems for online shopping. The method is based on extraction of colour-impression features from images based on images' colour features. The presented model is very useful, especially for domains where colour-based impressions play an important role in customer decisions for product selection and purchase, for example fashion and decoration domains. It is based on service-oriented architecture for greater flexibility and easier adaptation to different environments.

Other important multimedia and image analysis methods exist that can be applied to recommender systems, such as [40][41]. In our further work we shall extend the model with some of these methods in order to improve the model and extend the target domains by taking into consideration other information carried by images, such as combination of colour, shape and structure features, and other types of multimedia, such as video and sound.

ACKNOWLEDGMENT

We would like to thank The Matsumae International Foundation for their support and funding of this research.

REFERENCES

[1] G. Haubl 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.
 [2] Y. H. Cho, J. K. Kim, and S. H. Kim, "A personalized

recommender system based on web usage mining and decision tree induction," *Expert Systems with Applications*, vol. 23, no. 3, pp. 329-342, 2002.
 [3] J. B. Schafer, J. Konstan, and J. Riedl, "Recommender systems in e-commerce," in *Proceedings of the 1st ACM conference on Electronic commerce - EC '99*, pp. 158-166, 1999.
 [4] J. B. Schafer, J. A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications," *Data Min. Knowl. Discov.*, vol. 5, pp. 115-153, Jan. 2001.
 [5] E. Suh, S. Lim, H. Hwang, and S. Kim, "A prediction model for the purchase probability of anonymous customers to support real time web marketing: a case study," *Expert Systems with Applications*, vol. 27, no. 2, pp. 245-255, 2004.
 [6] R. D. Lawrence, G. S. Almasi, V. Kotlyar, M. S. Viveros, and S. S. Duri, "Personalization of Supermarket Product Recommendations," *Data Min. Knowl. Discov.*, vol. 5, pp. 11-32, Jan. 2001.
 [7] T. Y. Lee, S. Li, and R. Wei, "Needs-Centric Searching and Ranking Based on Customer Reviews," in *2008 10th IEEE Conference on E-Commerce Technology and the Fifth IEEE Conference on Enterprise Computing, E-Commerce and E-Services*, pp. 128-135, 2008.
 [8] O. Papaemmanouil, K. Pramataris, G. Prassas, and G. Doukidis, "A Recommender System for Online Shopping Based on Past Customer Behaviour," *BLED 2001 Proceedings*, Dec. 2001.
 [9] Junzhong Ji, Zhiqiang Sha, Chunnian Liu, and Ning Zhong, "Online recommendation based on customer shopping model in e-commerce," in *Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003)*, pp. 68-74.
 [10] J. K. Kim, Y. H. Cho, W. J. Kim, J. R. Kim, and J. H. Suh, "A personalized recommendation procedure for Internet shopping support," *Electronic Commerce Research and Applications*, vol. 1, no. 3, pp. 301-313, 2002.
 [11] Y. H. Cho and J. K. Kim, "Application of Web usage mining and product taxonomy to collaborative recommendations in e-commerce," *Expert Systems with Applications*, vol. 26, no. 2, pp. 233-246, 2004.
 [12] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, "Collaborative Filtering Recommender Systems," in *The adaptive web*, P. Brusilovsky, A. Kobsa, and W. Nejdl, Eds. Springer-Verlag Berlin, Heidelberg, 2007, pp. 291-324.
 [13] J. Li, Y. Xu, Y. Wang, and C. Chu, "Strongest Association Rules Mining for Personalized Recommendation," *Systems Engineering - Theory & Practice*, vol. 29, no. 8, pp. 144-152, 2009.
 [14] W. Yang, H. Cheng, and J. Dia, "A location-aware recommender system for mobile shopping environments," *Expert Systems with Applications*, vol. 34, no. 1, pp. 437-445, 2008.
 [15] L. Chen, F. Hsu, M. Chen, and Y. Hsu, "Developing recommender systems with the consideration of product profitability for sellers," *Information Sciences*, vol. 178, no. 4, pp. 1032-1048, 2008.
 [16] H. Wang and C. Wu, "A strategy-oriented operation module for recommender systems in E-commerce," *Computers & Operations Research*, 2010.
 [17] S. Huang, "Designing utility-based recommender systems for e-commerce: Evaluation of preference elicitation methods," *Electronic Commerce Research and Applications*, pp. -, 2010.
 [18] P. Symeonidis, A. Nanopoulos, A. Papadopoulos, and Y. Manolopoulos, "Collaborative recommender systems: Combining effectiveness and efficiency," *Expert Systems with Applications*, vol. 34, no. 4, pp. 2995-3013, 2008.
 [19] B. Jeong, J. Lee, and H. Cho, "An iterative semi-explicit rating method for building collaborative recommender systems," *Expert Systems with Applications*, vol. 36, no. 3, pp. 6181-6186, 2009.
 [20] P. Chou, P. Li, K. Chen, and M. Wu, "Integrating web mining and neural network for personalized e-commerce automatic service," *Expert Systems with Applications*, vol. 37, no. 4, pp. 2898-2910, 2010.
 [21] X. Zhang, J. Edwards, and J. Harding, "Personalised online sales using web usage data mining," *Computers in Industry*, vol. 58, no. 8, pp. 772-782, 2007.
 [22] A. Harada, *Report of modeling the evaluation structure of KANSEI*.

- University of Tsukuba, 1997.
- [23] Y. Kiyoki and X. Chen, "A semantic associative computation method for automatic decorative-multimedia creation with "Kansei" information," in *Proceedings of the Sixth Asia-Pacific Conference on Conceptual Modeling - Volume 96*, pp. 7–16, 2009.
- [24] S. Kurabayashi, T. Ueno, and Y. Kiyoki: "A Context-Based Whole Video Retrieval System with Dynamic Video Stream Analysis Mechanisms." In Proceedings of the 11th IEEE International Symposium on Multimedia (ISM2009), pp.505-510, San Diego, California, USA, December 14-16, 2009.
- [25] S. Kurabayashi and Y. Kiyoki, "MediaMatrix: A Video Stream Retrieval System with Mechanisms for Mining Contexts of Query Examples," in *Database Systems for Advanced Applications*, vol. 5982, H. Kitagawa, Y. Ishikawa, Q. Li, and C. Watanabe, Eds. Springer Berlin / Heidelberg, 2010, pp. 452-455.
- [26] S. Sasaki, Y. Itabashi, Y. Kiyoki, and X. Chen, "An Image-Query Creation Method for Representing Impression by Color-based Combination of Multiple Images," in *Proceeding of the 2009 conference on Information Modelling and Knowledge Bases XX*, pp. 105–112, 2009.
- [27] S. Kobayashi, *Color Image Scale*, 1st ed. Kodansha International, 1992.
- [28] A. H. Munsell, *Munsell Book of Color*. Baltimore: Munsell Color Company, 1929.
- [29] P. Valdez and A. Mehrabian, "Effects of color on emotions.," *Journal of Experimental Psychology*, vol. 123, no. 4, pp. 394-409, Dec. 1994.
- [30] S. Kobayashi, "The aim and method of the color image scale," *Color Research & Application*, vol. 6, no. 2, pp. 93-107, 1981.
- [31] Y. Kiyoki, T. Kitagawa, and T. Hayama, "A metadatabase system for semantic image search by a mathematical model of meaning," *SIGMOD Rec.*, vol. 23, pp. 34–41, Dec. 1994.
- [32] Y. Kiyoki, T. Kitagawa, and Y. Hitomi, "A fundamental framework for realizing semantic interoperability in a multidatabase environment," *Integr. Comput.-Aided Eng.*, vol. 2, pp. 3–20, Mar. 1995.
- [33] T. Kitagawa and Y. Kiyoki, "A mathematical model of meaning and its application to multidatabase systems," in *Third IEEE International Workshop on Research Issues in Data Engineering: Interoperability in Multidatabase Systems*, pp. 130 - 135, 1993.
- [34] M. B. Jurič, B. Mathew, and P. Sarang, *Business Process Execution Language for Web Services*, 2nd ed. Packt Publishing Ltd, 2006.
- [35] K. Pant and M. B. Jurič, *Business Process Driven SOA using BPMN and BPEL*. Packt Publishing Ltd, 2008.
- [36] T. Erl, *SOA Principles of Service Design*, 1st ed. Prentice Hall, 2007.
- [37] M. B. Jurič and M. Križevnik, *WS-BPEL 2.0 for SOA Composite Applications with Oracle SOA Suite 11g*. Packt Publishing Ltd, 2010.
- [38] OASIS, "Web Services Business Process Execution Language Version 2.0, OASIS Standard," <http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html>, 2007. [Online]. Available: <http://docs.oasis-open.org/wsbpel/2.0/wsbpel-v2.0.html>.
- [39] OMG, "Business Process Model and Notation (BPMN), Version 2.0 beta," 2009. [Online]. Available: <http://www.omg.org/cgi-bin/doc?dtc/09-08-14>. [Accessed: 20-Mar-2010].
- [40] A. R. Barakbah and Y. Kiyoki, "Image Search System with Automatic Weighting Mechanism for Selecting Features," in *The 6th International Conference on Information and Communication Technology and Systems*, 2010.
- [41] A. R. Barakbah and Y. Kiyoki, "An Image Search System with Analytical Functions for 3D Color Vector Quantization and Cluster-based Shape and Structure Features," in *Proceeding of the 2010 conference on Information Modelling and Knowledge Bases XXI*, pp. 169–187, 2010.

Sharing Emotional Information Using A Three Layer Model

Imen Tayari Meftah^{1,2} and Nhan Le Thanh¹
¹University of Nice Sophia Antipolis and CNRS

I3S Laboratory

Sophia Antipolis, France

Email: tayari@i3s.unice.fr, nhan.le-thanh@unice.fr

Chokri Ben Amar²

²University of Sfax and ENIS

REGIM laboratory

Sfax, Tunisia

Email: chokri.benamar@enis.rnu.tn

Abstract—In this study, we present a generic model to exchange emotional states information between heterogeneous multi-modal applications. Our proposal is composed of three distinct layers: the psychological layer, the formal computational layer and the language layer. The first layer represents the psychological theory adopted in our approach, The second layer is based on a formal multidimensional model. It matches with the psychological approach of the previous layer. The final layer uses XML to generate the final emotional data to be transferred through the network. The remainder of this article describes each layer of our model. The proposed model enables the exchange of the emotional states regardless to the modalities and sensors used in the detection step. Moreover our model permits to model not only the basic emotions (e.g., anger, sadness, fear) but also different types of complex emotions like simulated and masked emotions.

Keywords-emotion, multimodality, three layers model, Plutchik model, emotional exchanges, multidimensional spaces.

I. INTRODUCTION

The use of emotion in computers is becoming an increasingly important field for human-computer interaction. Indeed, affective computing is becoming a focus in interactive technological systems and more essential for communication, decision-making and behavior. There is a rising need for emotional state recognition in several domains, such as health monitoring, video games and human-computer interaction. The emotional information exchange between applications entails many problems such as application heterogeneity, complexity of emotional states, diversification of capture tools and dependence between treatment and physical sensors. The lack of a standard in human emotions modeling hinders the sharing of affective information between applications. As part of the research project Emotica, we define a generic model facilitating communication between heterogeneous multi-modal applications. Our proposal is composed of three distinct layers: the psychological layer, the formal computational layer and the language layer. This generic model is designed to be usable in a wide range of use cases, including modeling and representation of Emotional States. In this paper, we explain the role of each layer of our generic model. The remainder of this paper is organized as follows. In Section 2, we present the problem statement.

In Section 3, we describe the different parts of our model. Finally, we conclude in Section 4.

II. PROBLEM STATEMENT

Emotions are an important key component of human social interaction. Today there is a need for computers to understand this component in order to facilitate communication between users and to improve the credibility of interactions human-computer. Sharing emotional states becomes more and more important in human-machine interactive systems. Nevertheless, it entails many problems due to complexity of emotional states, diversification of capture tools and dependence between treatment and physical sensors. Emotion is a complex concept. Indeed there is no agreed model for the representation of emotional states. Many theories focused only on basic emotions [1]. Another ones introduced some complex emotions [2] but does not encompass all the emotional states. Moreover, in a natural setting, emotions can be manifested in many ways and, expressed and perceived through multiple modalities, such as facial expression, gesture, speech, or physiological variation. Each modality need a specific processing and use special techniques for the recognition step. The difficulty of sharing emotional information between many applications which use different modalities coming from the dependence between treatment and physical sensors and the lack of a standard human emotions modeling.

Current works on modeling and annotation of emotional states like Emotion Markup Language (EmotionML) [3] or Emotion Annotation and Representation Language (EARL) [4] aim to provide a standard for emotion exchange between applications, but they use natural languages to define emotions. They use words instead of concepts. For example, in EARL, joy would be represented by the following string "`<emotion category="joy" />`", which is the English word for the concept of joy and not the concept itself, which could be expressed in all languages (e.g., joie, farah, gioia). In this article, we propose a generic model, which can model any kind of complex emotion, and permits the exchange of emotional states between heterogeneous applications regardless to the modalities and sensors used in the detection step.

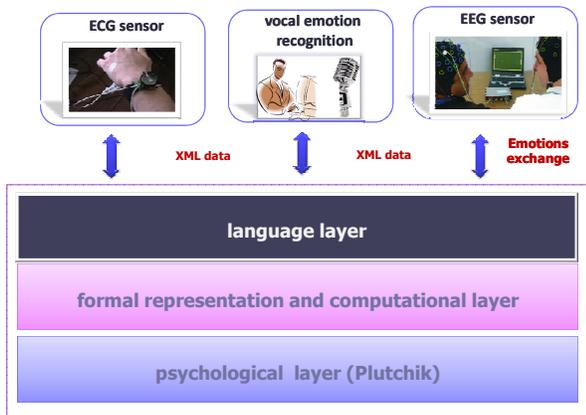


Figure 1. The Three-layer model

III. THE PROPOSED MODEL

Our approach is based on a hierarchical representation model composed by three distinct layers which are interdependent to ensure a maintenance of coherence of the model. Figure 1 shows a global schema of our proposed model. It is composed of three distinct layers: the psychological layer, the formal representation layer and the language layer.

A. The psychological layer

The psychological layer is the first layer of our model and it represents the psychological model that we chosen to represent the emotional state of users. Emotion is a complex concept. There is no consensus among psychological and linguistic theories on emotions. According to research in psychology, three major approaches to affect modeling can be distinguished [5]: dimensional, categorical, and appraisal-based approach. The dimensional approach models emotional properties in terms of emotion dimensions. It decomposes emotions over two orthogonal dimensions, namely arousal (from calm to excitement) and valence (from positive to negative) [6]. In the Appraisal theory, emotions are obtained from the subjective evaluations (appraisals) of events/stimulus that cause specific reactions in different people. Finally, categorical approach focus on identifying a small number of primary and distinct emotions. the number of basic emotions varies from one theory to the next: for instance, there are 6 basic emotions in the Fridja’s theory [7], 9 in the Tomkins’s theory [8] and 10 in the Izard’s theory [9]. Plutchik proposed a three-dimensional circumplex model (Figure 2) which describes the relationships between emotions. His model is very intuitive and easy including the idea that complex emotions are obtained by mixing primary ones. We opted for his approach as the basis of our model and will thus describe it in details.

1) *Plutchik model:* Robert Plutchik adopted a color metaphor for the combination of basic emotions [10]. He proposed a three-dimensional "circumplex model" (Figure

2), which describes the relationships between emotions. He argued for eight primary emotion arranged as four pairs of opposites: (Joy-Sadness, Fear-Anger, Surprise-Anticipation, Disgust-Trust) [10]. The vertical dimension represents intensity or level of arousal, and the circle represents degrees of similarity among the emotions. He suggested that non-basic emotions are obtained through the addition of basic emotions (color analogy, Plutchik, 1962) [11]. In his model, for instance, Love = Joy + trust and Delight = Surprise + Joy. Plutchik defined rules for building complex emotions out of basic ones. In practice, combination of emotions follows the method "dyads and triads" [12]. He defined the primary dyads emotions as the mixtures of two adjacent basic emotions. Secondary dyad includes emotions that are one step apart on the "emotion wheel", for instance Fear + Sadness = Despair. A tertiary emotion is generated from a mix of emotions that are two steps apart on the wheel (Surprise + Anger = Outrage).

In our work, we chose the Plutchik model to represent the psychological layer because it verifies many important conditions for the elaboration of our model and it explains emotions in terms of formulas that can be universally applied to all human beings [13]. First, the Plutchik model is based on 8 basic emotions encompassing the common five basic emotions. Then, it takes into account the intensity of emotion i.e., the level of arousal or the feeling degree of each basic emotion for example (terror, fear, apprehension). Finally, the Plutchik model is intuitive, very rich and it is the most complete model in literature because it permits to model complex emotions by using basic ones. Indeed, as we have seen, Plutchik defined the dyads and the triads which are combinations of basic emotions describing complex emotions which are regarded as emotions in usual life.

B. The formal computational layer

The formal computational layer is the second layer of our model. It matches the psychological approach of the first layer. It is the formal representation of Plutchik’s model and it is based on an algebraic representation using multidimensional vectors. In this layer, we represent every emotion as a vector in a space of 8 dimensions where every axis represents a basic emotion defined on the Plutchik theory .

First, we define our Base by $(B) = (\text{joy, sadness, trust, disgust, fear, anger, surprise, anticipation})$, which are the basic emotions on the Plutchik theory. So every emotion (e) can be expressed as a finite sum (called linear combination) of the basic elements.

$$(e) = \sum_{i=1}^8 \langle E, u_i \rangle u_i \tag{1}$$

Thus, $(e) = \alpha_1 \text{Joy} + \alpha_2 \text{sadness} + \alpha_3 \text{trust} + .. + \alpha_7 \text{surprise} + \alpha_8 \text{anticipation}$ where α_i are scalars and $u_i (i = 1..8)$ elements of the basis

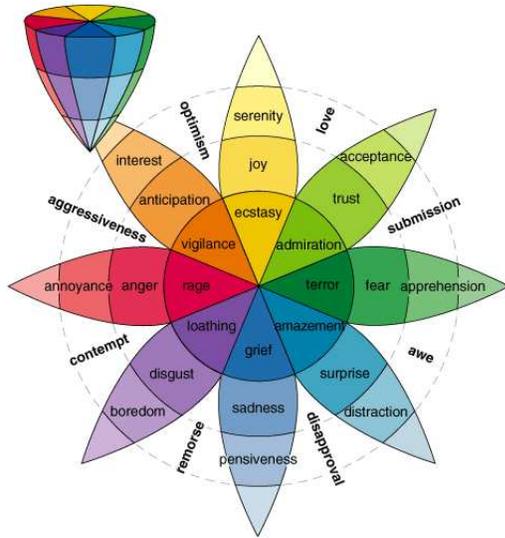


Figure 2. Plutchik's three-dimensional circumplex model

(B). Typically, the coordinates are represented as elements of a column vector E

$$E = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \alpha_8 \end{pmatrix}_B$$

where $\alpha_i \in [0,1]$ represent the intensity of the respective basic emotion. More the value of α_i gets nearer to 1, more the emotion is felt.

The proposed model takes into account the property of the intensity of the emotion. Indeed, each emotion can exist in varying degrees of intensity. The coefficients α_i determine the emotion intensity. According to the value of the coefficients α_i we can make the difference between annoyance, anger and rage or pleasure, joy and ecstasy. So, rage is the basic emotion anger with high intensity. The multidimensional representation of the formal computational layer provides the representation of an infinity of emotions and provides also a powerful mathematical tools for the analysis and the processing of these emotions. Indeed we can apply the usual basic algebraic operations on vectors like the addition, the scalar multiplication, the projection and the distance in an Euclidean space. We are going to detail only the addition, and the Euclidean distance. for more detail you can see [14].

1) *Vector addition:* We have seen in the previous paragraphs that the mixture of pairs of basic emotions resulted of complex emotion. Joy and trust for example produce the complex emotion "love". In this part we define the combination between emotions as the sum of two emotion vectors. This addition is defined as the maximum

Combinations & Opposites

„A mixture of any two primary emotions may be called a dyad.“

[often felt] PRIMARY DYADS	[sometimes felt] SECONDARY DYADS	[seldom felt] TERTIARY DYADS	OPPOSITES
joy + trust love	joy + fear guilt	joy + surprise delight	joy + sadness conflict
trust + fear submission	trust + surprise curiosity	trust + sadness sentimentality	trust + disgust conflict
fear + surprise alarm	fear + sadness despair	fear + disgust shame	fear + anger conflict
surprise + sadness disappointment	surprise + disgust ?	surprise + anger outrage	surprise + anticipation conflict
sadness + disgust remorse	sadness + anger envy	sadness + anticipation pessimism	
disgust + anger contempt	disgust + anticipation cynism	disgust + joy morbidness	
anger + anticipation aggression	anger + joy pride	anger + trust dominance	
anticipation + joy optimism	anticipation + trust fatalism	anticipation + fear anxiety	

Figure 3. Combination and opposites on the Plutchik's model

value of coefficients (term by term) [14]. Let E_{1u} and E_{2u} be two emotional vectors expressed in the basis (B) respectively by $(\lambda_1, \lambda_2, \dots, \lambda_8)$ and $(\lambda'_1, \lambda'_2, \dots, \lambda'_8)$. The addition of these two vectors is defined as:

$$E' = E_{1u} \oplus E_{2u} = \max(\lambda_i, \lambda'_i) \text{ for } 0 \leq i \leq 8 \quad (2)$$

In this sense, the vector representing the emotion love, which is mixture of joy and trust, is defined as:

$$E_{love} = E_{Joy} \oplus E_{trust} = \begin{pmatrix} \alpha_1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}_B \oplus \begin{pmatrix} 0 \\ 0 \\ \alpha_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}_B = \begin{pmatrix} \alpha_1 \\ 0 \\ \alpha_3 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}_B$$

where $\alpha_1 \neq 0$ et $\alpha_3 \neq 0$

In the same way we can obtain the "vector form" of the other complex emotions states defined by Plutchik. These emotions combinations are shown on Figure 3. Figure 4 shows an example of the using of the add operation on application of emotion detection. On this example the detection is done using two modalities. Each modality gives an emotion vector. The vector V_1 is given by the facial modality and the vector V_2 is given by the physiological modality. The final emotion vector V_f is given by the addition of this two vectors using equation 2.

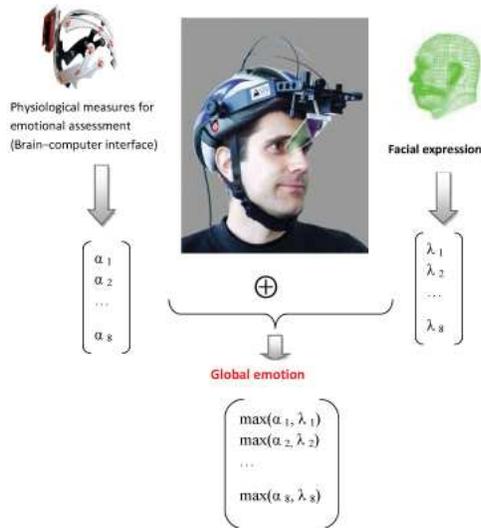


Figure 4. Multi-modality emotion recognition system

2) Euclidean distance (2-norm distance):

$$d(E, Y) = \sqrt{\sum_{i=1}^n (y_i - x_i)^2} \quad (3)$$

with $E \begin{pmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ x_n \end{pmatrix}_B$ and $Y \begin{pmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ y_n \end{pmatrix}_B$ are two vectors.

The proposed model is a continuous model providing the representation of infinity of emotions. Thus, to analyse a given vector and determine the nearest emotion from the known ones we need a tool to calculate the similitude from the vector and the known emotions. For this, we propose to use the Euclidean distance (2-norm distance) defined by equation 3. First, we have to generate a data base of emotions composed by the vectors of all emotions proposed by Plutchik given by Figure 2 and Figure 3. So, our emotion data base is composed by approximately 50 emotions and can be extended by others emotions. Then we have to compute for a given vector V1 the Euclidean distance between it and all the vectors of the data base. Finally we keep the vector of the data base minimizing this distance. This vector represents the nearest emotion of V1 and the computed distance gives an idea of the precision of this interpretation. For example, we can found that the nearest emotion for the vector V1 is "love" with a distance equals to zeros. We can affirm without doubts that V1 represents the emotion "love". More the distance from the nearest vector is important, less the interpretation is accurate. So the proposed method, using the Euclidean distance, permits to analyse automatically a given vector and provides the best

interpretation of this vector.

C. The language layer

The third layer of our model is the language layer. This layer provides encoding emotional information. We propose to use the eXtensible Mark-Up Language (XML) developed by the World Wide Web Consortium to annotate and represent emotional states of users.

XML, is a method for describing and encoding data. It is used for representation and transmission of data between application and organization. It is a text-based system meaning that both humans and machines can understand it directly, and is self-describing in so much as each data element can be traced to a definition [15]. For example, annotating a complex emotion detected using the voice modality (microphone) give the following XML structure:

```
<emotion>
<modality set= "basic-modality" >
<vector mode="voice">
<intensity axis="joy">0.8</intensity >
<intensity axis="sadness">0.0</intensity >
<intensity axis="trust">0.6</intensity >
<intensity axis="disgust">0.0</intensity >
<intensity axis="fear">0.0</intensity >
<intensity axis="anger">0.0</intensity >
<intensity axis="surprise">0.0</intensity >
<intensity axis="anticipation">0.0</intensity >
</vector>
</emotion>
```

The numeric values for the tag "intensity " indicate the intensity of the respective basic emotion going on, on a scale from 0 (emotion is not felt) to 1 (more the emotion is felt). Using the computational layer we can conclude that the felt emotion is a complex one because there is more than one axis with a value different from zeros. Moreover, using our algorithm based on the Euclidean distance, defined on the formal computational layer, we can conclude that the felt emotion is "love".

The next example was generated using two modalities: the heart rate modality and the facial expression modality. Each modality will give a vector with different coefficients.

```
<emotion>
<modality set= "multi-modality" count="2">
<vector mode="heart-rate">
<intensity axis="joy">0.0</intensity >
<intensity axis="sadness">0.4</intensity >
<intensity axis="trust">0.0</intensity >
<intensity axis="disgust">0.0</intensity >
<intensity axis="fear">0.8</intensity >
<intensity axis="anger">0.2</intensity >
```

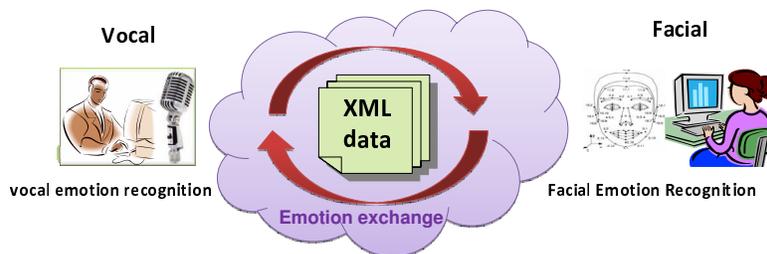


Figure 5. An example of sharing emotional states.

```

<intensity axis="surprise">0.0</intensity >
<intensity axis="anticipation">0.0</intensity >
</vector>
<vector mode="face">
<intensity axis="joy">0.0</intensity >
<intensity axis="sadness">0.7</intensity >
<intensity axis="trust">0.0</intensity >
<intensity axis="disgust">0.0</intensity >
<intensity axis="fear">0.1</intensity >
<intensity axis="anger">0.0</intensity >
<intensity axis="surprise">0.0</intensity >
<intensity axis="anticipation">0.0</intensity >
</vector>
</emotion>
    
```

In the last example, each modality gives a separate vector. Using the vector addition of the computational layer, we obtain the final vector describing the felt emotion

$$E = \begin{pmatrix} 0 \\ 0.7 \\ 0 \\ 0 \\ 0.8 \\ 0.2 \\ 0 \\ 0 \end{pmatrix}_B$$

The next step is to apply the distance algorithm to determine the most similar emotion of the data base to our vector. Figure 5 shows an example of exchanging emotional information between two users using our model. The first user (user1) uses a camera to detect emotions. The emotion detected using the facial expression is represented in XML data and sent to (user2). User2 analyses the xml data and uses our algorithm based on the Euclidean distance to find that the felt emotion by (user1) is "shame" and vice versa. we notice that (user2) can determine the emotion felt by (user1) regardless the modality used on the detection step and he can share emotional data with everyone without restriction. Moreover, this emotional information can be exploited by different application like 3D video games or serious games monitoring.

IV. CONCLUSION AND OPEN ISSUES

We have presented a generic model allowing the communication between various multi-modal applications. Our model is based on psychology research. It is composed of three distinct layers which are interdependent to ensure a maintenance of coherence of the model. The vectorial representation of emotions on the middle layer of our model allows powerful mathematical tools for the analysis and the processing of these emotions like addition (Figure 4), projection and decomposition [14]. This new model allows a transparent transfer of emotional states information between heterogeneous applications regardless to the modalities and sensors used in the detection step and provides the representation of infinity of emotions. For the future work, we would extend our model with adding another layer containing information relative to the user such as personality traits, social relations and emotional context.

REFERENCES

- [1] P. Ekman, *Basic emotions*, I. T. Dalgleish and T. Power, Eds. Sussex, U.K.: John Wiley & Sons, Ltd., 1999.
- [2] I. Albrecht, M. Schroder, J. Haber, and H.-P. Seidel, "Mixed feelings: expression of non-basic emotions in a muscle-based talking head," *Virtual Real.*, vol. 8, pp. 201–212, August 2005. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1086350.1086352>
- [3] P. Baggia et al., "Elements of an EmotionML 1.0," <http://www.w3.org/2005/Incubator/emotion/XGR-emotionml/>, Novembre 2008.
- [4] Humaine, "Humaine emotion annotation and representation language (earl): Proposal," <http://emotion-research.net/projects/humaine/earl/proposal>, June 2006.
- [5] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, pp. 484–495, 06/2008 2008.
- [6] J. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, no. 39, pp. 1161–1178, 1980.
- [7] N. H. Frijda, *The emotions*, Cambridge, Ed. Cambridge University Press, 1986.

- [8] S. Tomkins, "Affect as amplification: some modifications in theory," *Theories of emotions, vol. 1, New York, Academic Press.*, pp. 141–165, 1980.
- [9] C. E. Izard, *Human emotions*, S. Verlag, Ed. Plenum Press, New York, 1977.
- [10] R. Plutchik, *Emotion, a psychoevolutionary synthesis*. Harper and Row, New York, 1980.
- [11] ———, *The Emotions: Facts, Theory and a New Model*, ser. Studies in psychology. Random House, New York, 1962.
- [12] M. de Bonis, *Connaitre les émotions humaines*, Mardaga, Ed. Psychologie et sciences humaines, 1996, vol. 212.
- [13] T. Jonathan H and S. Jan E, *The Sociology of Emotions*. Cambridge University Press, 2005.
- [14] I. Tayari Meftah, N. L. Thanh, and C. Ben Amar, "Towards an algebraic modeling of emotional states," in *Internet and Web Applications and Services (ICIW), 2010 Fifth International Conference on*, May 2010, pp. 513 –518.
- [15] T. Bray, J. Paoli, C. M. Sperberg-Mcqueen, Eve, and F. Yergeau, Eds., *Extensible Markup Language (XML) 1.0*, 4th ed., ser. W3C Recommendation. W3C, August 2003. [Online]. Available: <http://www.w3.org/TR/REC-xml/>

An Evaluation of a Cluster-based Testbed for Peer-to-Peer Information Retrieval

Saloua Zammali

Dept. of Computer Science and Mathematics

Faculty of Sciences of Tunis

Tunis, Tunisia

Email: zammalisalwa@gmail.com

Khedija Arour

Dept. of Computer Science

National Institute of Applied Sciences and Technology of Tunisia

Tunis, Tunisia

Email: Khedija.arour@issatm.rnu.tn

Abstract—Peer-to-Peer (P2P) systems present an advantageous way to provide and share services [1]. Hence, P2P are the major technology of access upon various resources on Internet. Hence, P2P are the major technology of access upon various resources on Internet. A particularly intriguing class of distributed applications consists in Information Retrieval (IR) systems. The issue of Peer-to-Peer Information Retrieval (P2PIR) is being tackled by researchers attempting to provide valuable insights and to propose solutions to use it successfully. Nearly, all published studies have been evaluated by simulation means, using well known document collections (usually acquired from TREC). This practice leads to two problems: First, there is little justification in favor of the document distributions used by relevant studies and second, since different studies use different experimental testbeds, there is no common ground for comparing the solutions proposed. In this paper, we propose, CB a testbed for P2PIR based on P-Kmeans. CB, a cluster-based testbed, allows to distribute documents. This work marks the start of an effort to provide more realistic evaluation environments for P2PIR systems as well as to create a common ground to compare the current and future architectures.

Keywords-Testbed; P2P systems; Information retrieval.

I. INTRODUCTION

Peer-to-Peer (P2P) systems present an advantageous way to provide and share services [1]. Hence, P2P are the major technology of access upon various resources on Internet. Hence, P2P are the major technology of access upon various resources on Internet. A particularly intriguing class of distributed applications consists in Information Retrieval (IR) systems. The issue of Peer-to-Peer Information Retrieval (P2PIR) is being tackled by researchers attempting to provide valuable insights and to propose solutions to use it successfully. Nearly, all published studies have been evaluated by simulation means, using well known document collections (usually acquired from TREC). On the IR side, in a P2P network, the distribution of documents is, to a significant scale, a result of the previous location and retrieval. However, this also depends on the application specification and/or on other non-functional requirements that may be imposed (such as copyright considerations, etc.). Defining and simulating user behaviour, especially in a very large distributed system, is a complex and intimidating task. The problem with such approaches is a twofold. Firstly, there are cases where the documents distribution does not

successfully reflect the application scenario and therefore such evaluation results are hardly conclusive. Secondly, each individual considers a different testbed for experimental evaluation, the mutual comparison and the quantification of performance improvements become an impossible task.

Organising documents according to their content, and consequently, achieving more accurate and effective retrieval is, arguably, one of the principal goals of IR research. Document clustering has been a particularly active research field within the Information Retrieval (IR) community [2][3][4][5]. The reason behind this, apart from a natural human tendency [6], is that by clustering, documents relevant to the same topics tend to be grouped together (the Cluster Hypothesis [2]). Addressing these issues, we propose a testbed, suitable for the evaluation of P2PIR systems.

This paper is organized as follows. Section 2 defines the notion of testbed and Section 3 reviews related work about testbed in P2P retrieval. In Section 4, we detail our proposal and we are showing our first experimental results in Section 5. Section 6 concludes and gives some open issues.

II. NOTION OF TESTBED

- **Centralized Information Retrieval Testbed:**
Dekhtyar [7] defines IR testbed by the following formula:
Testbed = DataSet + Tasks + Answers + Evaluation measure + Data Formats. Indeed, a testbed must provide the documents and the queries to be raised on these documents. The answers to the queries are often data provided by experts, together with the relevance judgements. Evaluation measures are the tools which the testbed uses in order to test the relevance of the IR algorithms. Data Formats, relates to the existence of testbed under various formats of possible data.
- **Distributed Information Retrieval Testbed:**
In a distributed context, new information must be defined; how to distribute the data on the various nodes of a network and which replication law to apply? In addition, we define the elements which a distributed testbed must provide:
Distributed Testbed = documents collection+ queries collection + documents and queries distribution method

among peers + documents and queries replication method among peers + evaluation metrics+ queries responses.

Based on this notion of testbed, we propose in this paper, a cluster-based testbed. Before presenting the main features of our testbed, it is important to present a brief state of the art of some existing testbeds in a centralized and distributed context.

III. BACKGROUND AND RELATED WORK

A. Testbeds for centralized systems

For centralized Information Retrieval, there exist an important number of standard centralized benchmarks, such as the yearly competitions conducted by the Text Retrieval Conference, or TREC [8], DMOZ [9], etc.

TREC, co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program [10]. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval approach. In 2001 and 2002, the conference organized evaluating campaigns segmentation, indexing and searching content in the video [11]. For each version of TREC, NIST provides a collection of test. However, TREC is only available to registered participants of the conference. Other benchmarks repeatedly deployed in the literature include the Initiative for the Evaluation of XML Retrieval, or INEX [12], benchmark. The test collection consists of a set of XML documents, topics and relevance assessments. The topics and the relevance judgments are obtained through a collaborative effort from the participants. On the on-line topic submission, retrieval result submission, relevance judgment task, and evaluation metrics will be provided by INEX. Relevance assessments will be provided by the participating groups using INEXs on-line assessment system.

B. Testbeds for decentralized systems

In recent years, distributed information retrieval systems based on Peer-to-Peer (P2P) architectures have been increasingly attracting attention [13][14][15][16][17][18][19]. These systems usually consider a collaboration of peers where each peer stores a subset of the globally available documents. Being influenced by information retrieval in centralized systems, a substantial fraction of authors in the field of distributed IR evaluate their approaches by partitioning well known centralized IR testbed collections, such as the one provided by TREC, into (overlapping or disjoint) fragments. However, the assignment of documents to peers is not standardized and is performed differently by the authors, rendering the comparison of experimental results a bothersome task. Also, built testbeds is a challenge in distributed information retrieval systems and in particular in P2PIR systems.

IV. P-KMEANS: A CLUSTER-BASED TESTBED FOR P2PIR

In P2P network, each peer usually has a homogeneous collection of documents representing the interests of its user. Intuitively, clustering similar documents will help to discover useful resources and prune the searching space. Therefore, clustering similar documents will benefit information retrieval in P2P systems.

Clustering algorithms partition a set of objects, documents in our case, into groups called clusters. The classical algorithm K-means was introduced and drawing by Hartigan [20]. This algorithm is a classification tool that allows reserve a set of data in k homogeneous classes; k is fixed by the user. It affects each object, randomly, to a region and we iterate as follows: the centers of the different groups are recalculated and each object is assigned to a new group, based on the nearest center. Convergence is reached when the centers (also called (centroids)) are fixed [21].

But k-means also has drawbacks, among which we can mention: the method does not scale to large data collections. Indeed, most traditional methods of clustering are easily affordable but can not be applied to large collections of data. Their space complexity is often too great. It follows that it is interesting to seek an algorithm that is based on K-Means to enjoy these benefits and that adapts to a large-scale distributed environment. To obtain a semantics distribution on different network nodes, we first apply the algorithm K-Means on the document collection. We noticed that K-Means takes into account that small collections. Following this finding, we used an empirical study to determine the maximum number of documents processed by both K-Means. The results of this study is that this algorithm treats up to 5000 objects (*i.e.*, documents). For this, we thought about implementing a clustering algorithm (Peers KMeans or *P-KMeans*). The objective of this algorithm is to define a method of distributing documents on peer and overcome the main drawback of the algorithm K-Means. *P-KMeans* algorithm takes as input a documents collection \mathcal{DF} , the number of peers in the network k , the number of documents in \mathcal{DF} and the number of documents processed at each iteration. It will output the set of k -clusters \mathcal{P}_k .

The document distribution algorithm operates in three stages:

- Clustering of documents:
All documents in the \mathcal{DF} collection is partitioned according to the number of documents processed at each iteration n . The pseudo-code for the partitioning is given by Algorithm 2. The notations used are summarized in Table I. The partitioning algorithm takes a documents collection (*i.e.*, *documentsDF*), the number of documents in \mathcal{DF} and the number of documents processed at each iteration as input. It produces the subsets of documents (*i.e.*, *DocDef*). For any subset

k	:	peers number in network.
n	:	#documents processed at each iteration.
df_i	:	subset of documents.
\mathcal{N}	:	#documents in \mathcal{DF} .
\mathcal{DF}	:	documents collection.
\mathcal{P}_k	:	k -clusters.
$DocDef$:	set of all documents.
$clusterFiles$:	set of files containing documents clusters.
$centroidFiles$:	set of files containing centroids clusters.
$clusterCentroidsFiles$:	set of files containing centroids clusters.

Table I
P-KMeans ALGORITHM NOTATIONS.

df_i of $DocDef$, we apply *adaptedKMeans* algorithm (line 8-9) that takes df_i and the number of k peers in the network as input. It produces df_i documents groups (i.e., $clusterFiles$) and these centroids groups (i.e., $centroidFiles$).

- Clustering of centroids:
Centroids (i.e., $centroidFiles$) already generated previously, are grouped by K-Means algorithm (line 10) to produce centroids cluster ($clusterCentroidsFiles$).
- Mapping between document clustering and centroids clustering:

This step is the intermediate step between the clustering of documents and the clustering of centroids to obtain the conclusion of clustering documents. The pseudocode for this step is given by algorithm 3. The notations used are summarized in Table I. The mapping algorithm takes as input all documents clusters ($clusterFiles$) and all centroids clusters ($centroidFiles$), it produces k -clusters set (i.e., \mathcal{P}_k).

V. EXPERIMENTS

A. Experimental Environment

- PeerSim simulator:
To evaluate the approach proposed in this paper, we have chosen to use the PeerSim [22] simulator which is an open source tool written in Java. It has the advantage of being dedicated to the study of P2P systems. It has an open and modular architecture allowing it to be adapted to specific needs. More precisely we use an extension of PeerSim developed by the RARE project [23]. This extension can be seen as a PeerSim specialization for information retrieval.
- Source Data:
As a data set, we used "BigDataSet", produced under the RARE project [23]. It was obtained from a statistical analysis on data collected from the Gnutella [24] system and data TREC collection, which allows us to perform simulations in real conditions. BigDataSet is composed of a set of documents (25000), a queries set (4999), a set of peers (500) and a queries distribution on peers. It provides XML files describing the system

Algorithm 1: P-KMEANS

```

1 Algorithm: P-KMEANS( $\mathcal{DF}, k, \mathcal{N}, n$ )
Input:
 $\mathcal{DF}$ : documents collection.
 $k$ : peers number in network.
 $\mathcal{N}$ : documents number in  $\mathcal{DF}$ .
 $n$ : documents number at each iteration.
Output:
 $\mathcal{P}_k$ .
2 begin
3    $DocDef ::= \text{partitionDF}(\mathcal{DF}, \mathcal{N}, n)$ ;
4    $centroidFiles ::= \{\emptyset\}$ ;
5    $clusterFiles ::= \{\emptyset\}$ ;
6    $clusterCentroidsFiles ::= \{\emptyset\}$ ;
7   foreach  $df_i \in DocDef$  do
8      $clusterFiles ::= clusterFiles \cup$ 
9        $\text{adaptedKMeans}(df_i, k)$ ;
10     $centroidFiles ::= centroidFiles \cup$ 
11       $\text{adaptedKMeans}(df_i, k)$ ;
12     $clusterCentroidsFiles ::=$ 
13       $\text{KMeans}(centroidFiles, k)$ ;
14     $\mathcal{P}_k ::= \text{mapping}(clusterCentroidsFiles,$ 
15       $clusterFiles)$ ;
16  return ( $\mathcal{P}_k$ )
17 end
    
```

Algorithm 2: PARTITIONDF

```

1 Algorithm: partitionDF( $\mathcal{DF}, \mathcal{N}, n$ ) Input:
 $\mathcal{DF}$ : documents collection.
 $\mathcal{N}$ : documents number in  $\mathcal{DF}$ .
 $n$ : documents number traits at each iteration.
Output:
 $DocDef$ .
2 begin
3    $DocDef ::= \{\emptyset\}$ ;
4   for ( $i=0$ ;  $i \neq \mathcal{N}/(n-1)$ ;  $i++$ ) do
5      $df_i ::= \text{Partition}(\mathcal{DF}, n, i)$ ;
6      $DocDef ::= DocDef \cup df_i$ ;
7   return  $DocDef$ 
8 end
    
```

nodes and the documents they possess, as well as queries which will be launched on the network [25].

- Routing Algorithms
Routing models used here, are Gnutella and LPS. Gnutella is a system that used a simple constrained flooding approach for search. A query was forwarded to a fixed number of neighbors until its time-to-live (TTL) in terms of forwarding steps was exhausted or a loop was detected.

Algorithm 3: MAPPING

```

1 Algorithm: mapping(clusterCentroidsFiles,
  clusterFiles)
  Input:
  clusterCentroidsFiles: set of files containing
  centroids clusters.
  clusterFiles: set of files containing documents
  clusters.
  Output:
  DocDef.
2 begin
3   DocDef := {∅};
4   foreach (clusteri ⊂ clusterCentroidsFiles)
5     do
6       foreach (c ∈ clusteri) do
7         Dc := extractCentroidDocs(c,
          clusterFiles);
          c := Dc
8   return Pk;
9 end
    
```

LPS is an algorithm for routing queries based on learning implicit behavior of users that is deduced from queries history [26].

- Evaluation Metrics

In an IR system, the system's success or rejection is based on how effectiveness is measured. Recall (R), Precision (P) and F-score (the harmonic mean of precision and recall) measures have been widely used as fundamental measures to test the effectiveness of IR systems [27]. Let RDR , the number of relevant documents returned. Let RD , the number of relevant documents. Let DR , the number of documents returned. These measures are defined as follows:

$$R = \frac{RDR}{RD} \quad (1)$$

$$P = \frac{RDR}{DR} \quad (2)$$

$$F - score = 2 * \frac{P * R}{P + R} \quad (3)$$

- Initialize simulation parameters

The simulation, of both algorithms *LPS* and *Gnutella*, is based on the parameters:

- *TTL* (Time To Live): Maximum depth of research, initialized to 5.
- *Pmax*: Maximum number of peers which the query should be propagated to.
- *Overlay size*: Number of peers in the network, initialized to 500.
- *Replication degree*: We used the same *Zipf* replication degree that is 40.

B. Testbeds for Evaluating

We performed our evaluation using the testbeds proposed in [28]. These testbeds are based on "BigDataSet" collection, produced under the RARE project [23], and are designed to address a number of P2P IR applications through different document distributions and concentrations of relevant documents. The individual testbeds used are the following:

- **UBZR**: This testbed is designed for the simulation of systems where the documents are distributed uniformly across the peer population.
- **RBZR**: In this testbed, documents assignment is done in a completely random manner.
- **SB**: This testbed aims to reflect a P2PIR scenario. Relevant documents are distributed among a small number of peers. Each peer usually has a homogeneous collection of documents representing the interests of its user.

C. Experimental Results

Our experiments aim to determine the impact of different testbeds on routing algorithms performance. We compared *CB* (based on P-Kmeans algorithm) testbed with *UBZR*, *RBZR* and *SB*. Figures 1 and 3 show the results for Gnutella algorithm when applying the different testbeds. Figure 2 and 4 shows the results for LPS algorithm under different testbeds. Former tests presented here are, in our opinion, very encouraging. By comparing our testbed with existing ones, we evaluate that our testbed is competitive.

A search algorithm is substantially in influence by used type of distribution. Indeed, a semantics data distribution, such the case of *CB*, gives the best results compared to other testbeds. Indeed, distribute data according to thematic, such as *CB* testbed, sought may be beneficial both for flooding routing algorithm (case of Gnutella) and a semantic algorithm (case of LPS) and thus with recall and F-score.

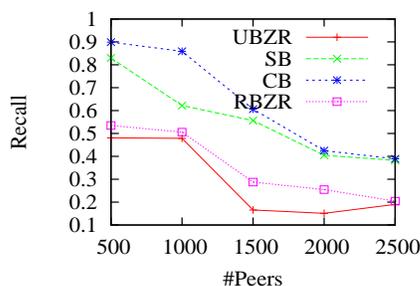


Figure 1. Relation between Recall and Nbr of peers according to different evaluation testbeds for Gnutella

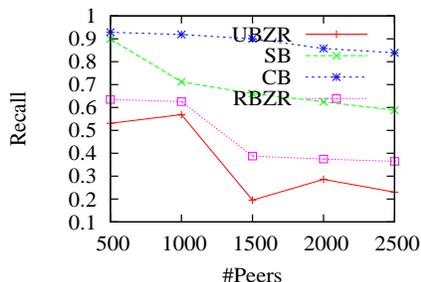


Figure 2. Relation between Recall and Nbr of peers according to different evaluation testbeds for LPS

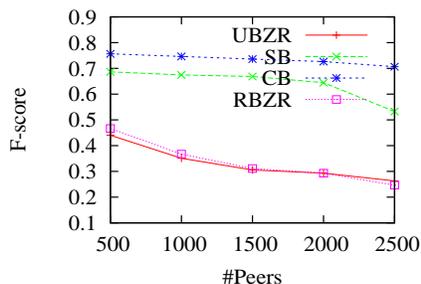


Figure 3. Relation between F-score and Nbr of peers according to different evaluation testbeds for Gnutella

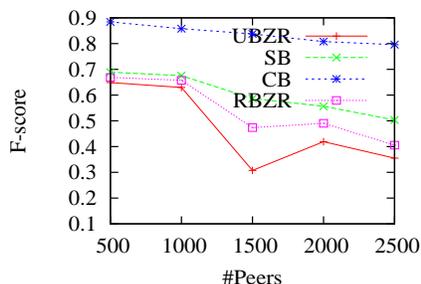


Figure 4. Relation between F-score and Nbr of peers according to different evaluation testbeds for LPS

VI. CONCLUSION AND FUTURE WORKS

The field of information retrieval is very experimental in nature. We identify the need to create testbeds for information retrieval experimentation. We propose *CB*, a testbed for P2PIR, based on clustering algorithm (P-Kmeans). In our testbed, recall and F-score (harmonic mean) are implemented as two instances of the evaluation element. Finally, this work can be followed by the use of other collections (e.g. INEX, DMOZ, etc.) and development of more realistic distribution methods, by building a real centralized collection (documents, queries and relevance judgments) from P2P network data.

REFERENCES

- [1] D. L. Lee, D. J. Zhao, and Q. Luo, "Information retrieval in a peer-to-peer environment," in *Proceedings of the 1st international conference on Scalable information systems*, New York, NY, USA, 2006.
- [2] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.
- [3] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of American Statistical Association*, vol. 58, no. 301, pp. 236–244, March 1963.
- [4] F. Murtagh, "A survey of recent advances in hierarchical clustering algorithms," *The Computer Journal*, vol. 26, no. 4, pp. 354–359, 1983.
- [5] Q. He, "A review of clustering algorithms as applied in IR," Ph.D. dissertation, University of Illinois, 1999.
- [6] R. M. Cormack, "A review of classification," *Journal of the Royal Statistical Society*, vol. Series A, no. 134, pp. 321–353, 1971.
- [7] A. Dekhtyar and J. Hayes, "Good Benchmarks are Hard To Find: Toward the Benchmark for Information Retrieval Applications in Software Engineering," in *IEEE International Conference on Software Maintenance (ICSM 2007)*, France, October 2007, pp. 1–3.
- [8] TREC, "Trec web site," <http://trec.nist.gov/>, 18/February, 2010.
- [9] DMOZ, "Dmoz web site," <http://www.dmoz.org/>, 18/February, 2010.
- [10] D. K. Harman, "Overview of the first text retrieval conference (trec1)," in *Proceedings of the First Text REtrieval Conference (TREC1)*, NIST Special Publication, 1993, pp. 207–500.
- [11] S. Robertson, "Introduction to the special issue: Overview of the TREC routing and filtering tasks," *Information Retrieval*, vol. 5, no. 2-3, pp. 127–137, 2002.
- [12] INEX, "Inex web site," <http://inex.is.informatik.uni-duisburg.de/>, 18/February, 2010.
- [13] J. Lu and J. Callan, "Content-based retrieval in hybrid peer-to-peer networks," in *Conference on Information and knowledge management (CIKM 2003)*, New Orleans, USA, 2003, pp. 199–206.
- [14] K. Aberer, P. Cudr-Mauroux, M. Hauswirth, and T. V. Gridvine, "Pelt: Building internet-scale semantic overlay networks," in *International Semantic Web Conference (ISWC 2004)*, 2004.
- [15] S. Idreos, M. Koubarakis, and C. Tryfonopoulos, "P2p-diet: An extensible p2p service that unifies ad-hoc and continuous querying in super-peer networks," in *Special Interest Group on Management of Data (SIGMOD 2004)*, 2004.
- [16] H. Nottelmann, G. Fischer, A. Titarenko, and A. Nurzenski, "An integrated approach for searching and browsing in heterogeneous peer-to-peer networks," in *Heterogeneous and Distributed Information Retrieval (HDIR 2005)*, 2005.

- [17] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer, "Minerva: Collaborative p2p search," in *Very Large Data Bases (VLDB 2005)*, 2005, pp. 1263–1266.
- [18] F. C. Acuna, C. Peery, R. P. Martin, and T. D. Nguyen, "Planetp: Using gossiping to build content addressable peer-to-peer information sharing communities," 2003.
- [19] T. Suel, C. Mathur, J. wen Wu, J. Zhang, A. Delis, M. Kharrazi, X. Long, and K. Shanmugasundaram, "Odyssey: A peer-to-peer architecture for scalable web search and information retrieval," in *Web and Databases (WebDB 2003)*, 2003.
- [20] H. Bock, "Origins and extensions of the k-means algorithm in cluster analysis," *Electronic Journal for History of Probability and Statistics*, vol. 4, no. 2, pp. 1–18, December 2008.
- [21] S. Siersdorfer and S. Sizov, "Automatic document organization in a p2p environment," in *European Conference on Information Retrieval (ECIR 2006)*, London, 2006, pp. 265–276.
- [22] Peersim, "The peersim simulator," <http://peersim.sf.net>, 18/February, 2010.
- [23] RARE, "Le projet rare (routage optimisé par apprentissage de requêtes)," in <http://www-inf.int-evry.fr/defude/RARE/>, 2010.
- [24] Gnutella, "Gnutella web site," <http://www.gnutella.com/>, 18/February, 2010.
- [25] R. Mghirbi, K. Arour, Y. Slimani, and B. Defude, "A Profile-Based Aggregation Model in a Peer-To-Peer Information Retrieval System," in *Data Management in Grid and P2P Systems (Globe 2010)*, Spain, September 2010, pp. 148–159.
- [26] T. Yeferny and K. Arour, "LearningPeerSelection: A Query Routing Approach for Information Retrieval in P2P systems," in *International Conference on Internet and Web Applications and Services (ICIW 2010)*, Spain, May 2010, pp. 235–241.
- [27] M. Renda and U. Straccia, "Web metasearch: rank vs. score based rank aggregation methods," in *Symposium on Applied computing (SAC 2003)*, New York, 2003, pp. 841–846.
- [28] S. Zammali and K. Arour, "P2PIRB: Benchmarking Framework for P2PIR," in *Data Management in Grid and P2P Systems (Globe 2010)*, Spain, September 2010, pp. 100–111.

Reviving the Innovative Process of Design Thinking

Justus Broß*, Christine Noweski† and Christoph Meinel‡
Hasso Plattner Institute at the University of Potsdam
Prof.-Dr.-Helmert-Strasse 2-3, 14482 Potsdam, Germany
Fax: +49 (0)331 5509-325

*Telephone: +49 (0)331 5509-524, Email: justus.bross@hpi.uni-potsdam.de

†Telephone: +49 (0)331 5509-535, Email: christine.noweski@hpi.uni-potsdam.de

‡Telephone: +49 (0)331 5509-222, Email: office-meinel@hpi.uni-potsdam.de

Abstract—The application of weblogs in educational environments is enjoying an increasingly good reputation. In this paper, we describe our experiences with weblogs, supporting the innovation seeking process 'design thinking' that has recently received attention in various fields of interest. In face of this broad range of 'design thinking' usage, it was surprising to us, how little technology is employed in the corresponding teamwork processes. In this paper, we therefore stress the idea that weblogs may greatly support design thinking teams. Firstly, by enabling and supporting the formation of team communities across barriers of individual diversity, space and time, and secondly by supporting the process of design thinking itself.

Keywords-Web-based Collaboration; Management; Documentation; Performance; Design; Experimentation; Human Factors.

I. INTRODUCTION

Nowadays, no individual alone could ever know all that there is to know [1]. To remain competitive in the 21st- century global economy, knowledge worker must be increasingly specialized, and at the same time cooperate in diverse setup teams.

Diversity has been credited with myriad positive outcomes for team performance. Research, meanwhile has shown that the performance advantages of diverse setup teamwork are often found under very narrow conditions [2]. Both, experimental and field researches show that teams often do not reach their potential [3]. When groups collaborate there may be a tendency to loaf, to prematurely evaluate group products, or for some individuals to dominate the group process or distract the group from its goals.

This creates a demand for sophisticated coordination and management [4]. Techniques that assure efficient interaction, appropriate leadership, and motivating goals shall help groups overcome some of the negative forces. Design thinking offers such a broad set of techniques and opens up a space where creative teamwork can lead to innovation.

In this paper, we stress the idea that weblogs as such a technique among others may greatly support the innovative process of Design Thinking. Firstly, by enabling the formation of teams and communities across barriers of individual diversity, space and time, and secondly by supporting the process of design thinking itself. This paper underpins this

argumentation with the subsequent arrangement of sections: The subsequent section introduces the general concept of Design Thinking. Section three elaborates upon the question to what extent modern information and communication technology might enhance the process of Design Thinking. In doing so it delves into the question to what extent weblogs might be a general value-add in the teaching and learning environment and particularly in the process of Design Thinking. The following use case in section four is about the implementation of a weblog in a highly dynamic and innovative learning- and teaching- environment. It will indicate that weblogs - if implemented correctly in a specific context of application - can indeed greatly improve the process of Design Thinking. We finally discuss terms of success and give an outlook on further research questions.

II. WHAT IS DESIGN THINKING?

Recently, the term 'design thinking' has received attention in various fields of interest. The concept has its roots in research on how designers comprise wicked problems [5] and develop novel and viable solutions. Originally investigated in domains like architecture and industrial design, the initial research focuses on cognitive models supporting the generation, condensation, and creative transformation of design knowledge and concepts [6] [7] [8]. Building on that, design thinking was further developed and translated into metadisciplinary frameworks detached from designers' professional domains and was applied to various disciplines and fields of innovation. Design agencies such as IDEO promote working methods labelled with this term and inspire large scale companies like Procter & Gamble and SAP to design thinking' approaches to innovation [1]. As such, it supports the useful exchange of knowledge which has shown to be crucial for innovation processes. Nowadays, the term has therefore expanded into academic curricula beyond traditional design programs, as, for instance, at Rotman School of Management (Toronto) in the context of MBA education, and at the d.Schools in Stanford and Potsdam which offer design thinking education specifically to non-designers [9] [10]. Here, post-graduate students learn to work across their highly specialized particular disciplines in

diverse teams. This diversity includes cognitive, disciplinary and social diversity. To enable the analysis and processing of a wide scope of challenges and to deal with so-called "wicked problems" [11], teams comprising members with backgrounds in distinct disciplines are required. While the team setup is interdisciplinary, the participants interact in an open dialogue that transcends their respective disciplines, accepting each perspective as of equal importance and relating the different perspectives to each other. This necessitates exchange between domain languages and between everyday practices of different fields. In order to facilitate this approach, student teams are attended by teachers with the competence of moderation, mediation, association and transfer needed for this interdisciplinary mode of working [12].

One of the rather well proven success metrics of design thinking projects is the alternating use of divergent (where the questioner attempts to diverge from facts to the possibilities that can be created from them - concept domain) and convergent (where the questioner attempts to converge on and reveal "facts"- knowledge domain) thinking which is actively promoted through frequent feedback and testing. This change of thinking style results in a specific creative process, open to iteration. Song et al. [13] show that teams cycling between divergent and convergent patterns of thinking and questioning perform better than teams that have little variation over the design process. A frequent shift furthermore promotes the pooling of unshared information [14] [15], another established indicator for successful teamwork, which is the basis of what is often described by 'creating something larger than the sum of the individual input'.

The optimal team setup for Design Thinking teamwork projects is diverse: interdisciplinarity, mixed social backgrounds and cognitive diversity are some of the keywords here. Design thinking allows these diverse setup teams to develop a mutual understanding due to its strong emphasis on team-based learning regarding both the problem and its potential solutions.

Therefore, design thinking uses a broad variety of instruments. In addition to the predominant use of whiteboards, post-its and simple pens, digital documentation and communication applications are employed as well. Most of the time, a combination of analogue and digital instruments will be reality.

III. SUPPORTING HIGH DIVERSITY TEAMWORK

Until this decade, the ability to use technology to enable networked innovation was very limited. The primary technologies used to facilitate group innovation were paper and, more recently, the whiteboard and dry erase marker. Since then, a great deal has happened in the past decade that is revolutionizing collaborative innovation. New communication and collaboration platforms, media, and tools now allow many-to-many collaboration at a scale and cost that

could never have been achieved in the past. The Internet, an overnight success three decades in the making, along with its younger cousin the Web, really does change everything. For the first time, we now have tools that enable the free exchange of information across many individuals with remarkably low friction. Unfortunately, by seeking the rare brilliance of a limited few instead of the statistically likely success of the connected many, the "lone genius" worldview has limited our ability to make meaningful progress in everything from technology, to organizations, to education, and all the way to society [16]. We have done very little to systematically develop technology to support the innovation process. Overall, we are still in the "horseless carriage" days of living in a truly networked world. We can do better, but how do we begin to engage this new way of being? We believe a path to the future can be found by paying conscious attention to evidence of what works in the world today, and by asking: What are some of the enabling collaborative tools available today? There are many web-enabled collaborative tools that can be used for innovation seeking teamwork:

- *Instant messaging*: The ability to easily send short messages back and forth to others who are present using computers and mobile devices.
- *Conference calling*: Previously only available to corporate entities, now virtually anyone with a connected computer can initiate and participate in a conference call with others worldwide.
- *Video conferencing*: This is the addition of live video to conference calls or one-to-one messaging.
- *Shared whiteboards and documents*: These allow people to interact in real time and share documents, photos, drawings or presentations where anyone can edit or annotate the shared media. It reinforces collaboration and iteration.
- *Virtual spaces*: Web offerings to interact in real time within a virtual three-dimensional world.
- *Question and answer sites / Portals*: Many websites allow groups of people to easily share their knowledge and create new value.
- *Wikis and weblogs*: Web tools that have become widely available in recent years, making publishing quick and easy. They encourage dialogue and sharing, via asynchronous posting of comments, documents, discussions, and editing of shared media. Blog search engines, such as Google Blog Search and Technorati, BlogPulse or BlogIntelligence [17], allow people to easily search a huge quantity of very dynamic information. By transcending time, space and language barriers, blogs thus enable the exchange of knowledge across conventional borders [18] [19] [20].

A. Weblogs in creative teaching and learning environment

The focus of this paper is on this last group of web-enabled tools, since we believe that weblogs can spur the

process of Design Thinking most [21]. Technically, weblogs are an easy-to-use, web-enabled Content Management System (CMS), in which dated articles ("postings") as well as comments on these posts are displayed in reverse chronological order. Stephen Downes for instance postulates to define weblogs independently of their content: "Blogging is something defined by format and process, not by content" [22]. Blogs are therefore a form of micropublishing that are at first undefined regarding their field of application [23]. This undefined point of origin makes them so flexible for numerous other potential purposes, beginning with personal diaries, reaching over to knowledge- and activity management platforms, and finally to content-related and journalistic web offerings. Weblogs belong to the group of "social software": simple, easy-to-use, and flexible applications that not just enable, but also facilitate cooperative gathering of content. The consent among all currently existing social software tools next to weblogs, is that the surplus value is generated out of collaborative (social) activity. Social software therefore enables affiliation to (social) networks, as well as structuring and channeling of attention towards a certain field of interest [21].

Research on computer-based or electronic brainstorming has found that electronic groups can perform about as well as or better than nominal groups [24]. With the electronic technology, group members can share ideas simultaneously, be anonymous to other group members (low evaluation apprehension), and be accountable for their individual performance on their station (low social loafing). On top of that, insuring individual accountability can enhance performance in groups [25]. Kayser [26] stresses the similar experimental finding, that the use of a written or electronic exchange process is one important factor to enable groups to reach a high level of creativity. This is due to the fact that team sessions should be followed by individual idea-generation sessions to fully tap the cognitive benefits of an exchange process. The individual part structures the process to minimize production blocking, evaluation apprehension and social loafing.

The application of weblogs in teaching or learning environments is as a consequence enjoying an increasingly good reputation. Diverse authors attribute blogs the potential to be a transformational and innovation-enhancing technology in this regard [23]. Summarizing the findings of their qualitative study, Efimova and Fiedler [27], for instance, pool the benefits of the technology through the following characteristics:

- The representation of multiple perspectives
- Revelation of synergies of individual and collaborative learning,
- The acquisition of meta-learning-strategies as well as
- The facilitation of access to experts.

It seems to be a competitive advantage when students prepare, document and allocate their work, questions, sug-

gestions or recommendations in form of postings in a quick and easy to handle ready-made manner [21]. Highly ranked institutions such as the Harvard Law School start to strategically implement weblogs as digital portfolios for their student- and teaching body [23].

The core of a blog is a so-called 'personal learning space' in which personal artifacts are largely shared with a non-uniform group of audiences. The individual process of students is hereby documented and can at all times be reflected and reused by themselves and all other potential stakeholder. This space quickly becomes a complex ePortfolio [28], defined as "[...] a Web-based information management system that uses electronic media and services to enable learners to build and maintain a digital repository of artifacts for demonstration of competence and reflection on their learning" [29].

B. Weblogs – a Design Thinking supporting technology

The usage of a weblog as central and strategic point of information and communication within a design thinking process offers not only the integration of all information collected via these tools, it supports furthermore an active division between collaborative and individual working phases. This is crucial for a successful integration of as diverse knowledge and ideas as possible. There is a misbelief, that teamwork means working in a group at all times of a project. It has been proven to be wrong [30]. Instead, it is crucial for the success of teamwork, to shift between face to face group interaction and individual action. This allows all team members to work at their speed and to put forward their ideas without being interrupted or rated before an idea is fully elaborated. It also helps the team to grow together and to value each member as an individual, if they have the chance to put forward their knowledge, ideas and remarks in individual sessions.

This moment of true teambuilding is what most design thinking methods are designed for: brainstorming, storytelling, user research, prototyping and many more are neither new nor unique. Their power lies in the right combination and application over a creative thinking process. All these methods support the active solidarization of the individuals into one team, following the similarity-attraction paradigm. According to Mannix and Neale [31], the predictions of this paradigm are straightforward: Similarity on attributes such as attitudes, values and beliefs and behavior facilitate interpersonal attraction and liking - basic needs for a trust- and successful collaboration.

IV. USE-CASE D-SCHOOL BLOG

The following use case is about the implementation of a weblog in a highly dynamic, modern and innovative learning- and teaching- environment that focuses on the innovational culture of "Design Thinking". It will indicate that weblogs - if implemented correctly in a specific context

of application - can indeed improve the traditional and conservative way of education [21]. Our use case supports Ojala's reasoning that blogs give room to alternative views and opinions and inaugurate a distinct culture of thinking outside the mainstream [18].

A. Motivation and Background

The initial motivation to start the D-School-blog-project was to support the so-called "Innovation Lab", set up by the Hasso Plattner Design Institutes of Potsdam (Germany) and Stanford (USA) at the world's biggest IT fair CeBIT in Hannover in 2009 [32].

On each of the seven fair days, the lab used a predefined design thinking process, to develop original ideas for daily challenges with the overall goal to 'humanize' IT. CeBIT visitors, as well as the American and German students and professors from their respective locations in Stanford and at the fairground were joining forces in big, creative and interdisciplinary team in these 24-hour projects to develop fresh ideas for user-friendly products and services. The overall goal of the 'Innovation Lab' was to turn CeBIT into an opportunity for every fair visitor to experience innovation first-hand and to show that innovation can be taught and learned.

Communication of preliminary results within the distributed team was supported by the employment of free tools and web services such as SKYPE, YOUTUBE, PICASA WEB GALLERIES and more. Anyone with Internet access was able to follow, comment and contribute to the progress of the Innovation Lab through the 'D-School-Blog'.

B. Design, Structure and Content

For creative endeavours that require composition of novel artifacts, enhanced interfaces shall facilitate the exploration of alternatives, prevent unproductive choices, and enable easy backtracking. Therefore the interface is of crucial importance. Another requirement is to find acceptance among a very broad user range, as this is a crucial aspect of design thinking teams. Wide spreads of technology and premature knowledge and willingness to explore the digital space is one of the difficulties one faces when building a weblog.

To facilitate the blog's acceptance inside the DT community, its graphical presentation was realized along the Corporate Design of the D-School in Potsdam. It displays a sense of playfulness as well as an uncommon, creative character that is well-known and closely associated with Design Thinking. The starting page of the D-School-Blog was therefore subdivided into five major areas:

Area "A" is comparable to the typical header of any weblog, including search functionality, multi-language support, information on why the blog was initiated and most notably the blog's navigational area that links to the following four main categories:

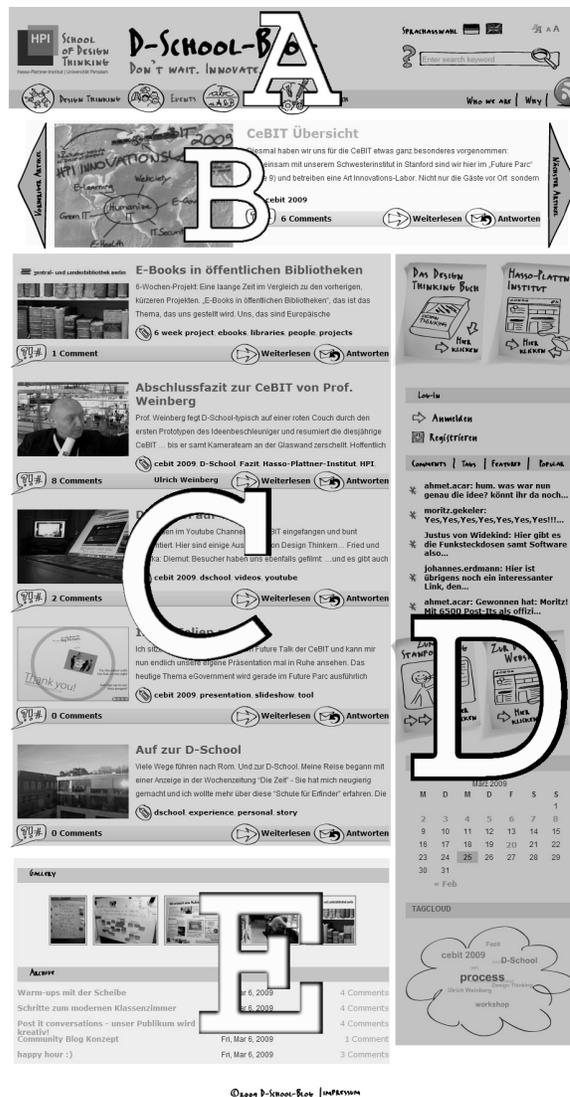


Figure 1. D-School Blog Starting Page - accessible via <https://d-school-blog.hpi-web.de/>

- The 'Design Thinking' category includes all kinds of posts that generally describe the innovational process of Design Thinking and that cannot be included in the following more specific categories.
- "Events" comprise postings that link Design Thinking with specific events like the above-mentioned innovation lab at the CeBIT, the presentation of the first Design Thinking book and others.
- "Classes" specifically focuses on the Design projects undertaken in the Design School of the HPI in Potsdam.
- The "Research" category is exclusively reserved for all work related to the bilateral Design Thinking Research program of the HPI and the Stanford University.

Label "B" in Figure 1 displays the featured articles in the D-Blog. This plugin-enabled functionality adds a number of

features to the standard blogging software of WORDPRESS that allow moderators to easily write and organize series of posts and display that series dynamically in the blog. This feature allows drawing extra attention towards a specific event, project or topic regarding Design Thinking that is being covered by multiple posts. In the case at hand, we used this area within the blog to display the different Design Challenges called out during our "innovation Lab" at the CeBIT in the most prominent way possible.

Any article posted in the D-blog finds its way into the main content area labeled "C". Usually, the posts are displayed in reverse chronological order, typical for weblogs. The sidebar of the D-blog is presented in a weblog-common way in the area labeled "D". Here, the blog visitors find important Design Thinking-related links, fields for registration and log-on, as well as a mask displaying the latest comments posted and the most popular posts written, as well as a list of the featured articles as also displayed in area "B". Here, you also find the list of tags assigned to the single posts, also displayed in the "tagcloud" at the bottom of the sidebar.

Area "E" is the last major division of the blog's starting page. Since multimedia content is generated in masses within the DT-process, a corresponding preview-gallery also needed to be placed on the starting page. The archive with "out-dated" posts finds its place at the very end of "area" E, only followed by copyright information and imprint.

Bross et al. [21] subdivide the D-School-Blog's community into the following user groups:

- Scientific staff, students and alumni of the D-School in Potsdam and in Stanford,
- the editorial staff of the D-blog that was also responsible for its technical realization,
- people that are interested in the Topic of Design Thinking and that are willing to contributing to the process by writing their own articles or commenting on the content of user-groups one till three.

The diverse background of those involved in this project obliged the editorial staff to provide technical support to parts of the community by giving continuous support of how to use a medium such as weblogs.

C. Proof of concept

The success of the community blogging concept is indicated by the usage statistics of the D-School-Blog. The initial pool of users consisted of students and alumni taking part in the 24h challenge and of design thinking researchers interested in participation at the CeBIT in 2009. A total of 55 user accounts have been set up for the launch of the blog, of which 39 have logged in within the first three days. Nearly 2 years later, the blog matured from a single-event documentation tool to a Design Thinking community platform that is fully integrated into the curriculum of the D-School in Potsdam. It is also prominently linked from the D-School's homepage (see [33]). 140 different authors

are by now regularly publishing new posting on the weblog. In total, the D-School-Blog so far welcomed in excess of 27.000 different users on its pages, who generated more than 100.000 hits. We also believe that the D-School-Blog is increasingly attracting interest in and outside the virtual borders of Germany. Indicators for this assumption are close to 25.000 referrers regarding the d-school-blog from external web pages, as well as the rising number of more than 300 followers on the accompanying twitter feed (twitter.com/dschool_potsdam). More to the point, selected projects of the D-School that have been documented in its blog, such as BRING.BUDDY (see [34]), have made it into a German newspaper with a wide circulation [35] and even into a popular German television show [36].

D. Critical success factors

Blogs may support the team building process but meanwhile need to find acceptance among all stakeholder. This is a tough task, due to the diverse setup that we have described earlier on. None of the less, we believe that if the following rules are abided, a blog should not be a source of friction. The user should therefore be allowed

- to take an holistic view of the source data or raw material with which they work
- to suspend judgment on any matter at any time and be able to return to that suspended state easily
- to be able to make unplanned deviations; return to old ideas and goals, formulate, as well as solve, problems and
- to re-formulate the problem space as their understanding of the domain or state of the problem changes.

Despite of the numerous advantages weblogs might incorporate for an innovative environment such as the Idea of Design Thinking, there is a downside in their application as well. Several experiments show, that acquiring knowledge of others through social software (e.g. weblogs) without personal interaction cannot fully replace the depth of understanding of face-to-face interaction (e.g. through non-verbal communication like mimic or gesture) [37]. We therefore strongly support the combination of real "physical" networking and virtual networking in order to leverage social software to the maximum.

Nardi et al. [38] state on this behalf that it is not what you know - it is who you know in the modern world that is most important in helping you getting a job or task done satisfactorily.

In other words, social networking increases the resources that can be leveraged through interpersonal relationships - thus social capital [37] [39]. Scholars transcribe professional networking with maintaining contacts, socializing, engaging in professional activities such as attending conferences, participating in community groups, and increasing visibility to others [40]. It thus equally includes emailing, participating

in social networks such as FACEBOOK and using weblogs in the modern era of social media [29].

V. SUGGESTIONS FOR FURTHER RESEARCH

Creativity is a socially defined activity. As such, measures of a creativity support tool's success are partially dependent on how success is defined and evaluated within a specific community of practice. Consequently, traditional measures such as performance or efficiency, while still important, are only one lens with which to view the value of a creativity support tool. To gain a more holistic perspective of how a tool influences the creative process, one may find it necessary to define new ways of measuring the impact of a creativity support tool on the problem solving process, where these metrics are derived from practices deemed important by the community under investigation. The following types of research questions that should be asked in evaluation studies of blogs supporting design thinking teams:

- Is this technique better than existing practice (Post-Its, Whiteboard, etc.)?
- Does it expand its use to other contexts?
- Have you learned how to improve this tool based on this evaluation?
- How does the tool/technique influence the creative process?
- What facets of creativity are affected and to what degree?
- How brittle is the tool/technique?
- How accepted is it by the users over the long term?
- Does it celebrate diversity?
- How does this method complement others in the family of tools/techniques?
- What is the task-to-technology "fit"?

VI. CONCLUSION

On top of the general advantages that weblogs might have in a teaching or learning environment (refer to section 3.1), we identified several characteristics that are specifically useful when deployed in the context of Design Thinking. One crucial success factor is their ability to actively promote and support frequent feedback and testing of their rationale. This allows users to fundamentally change their style of thinking and make room for specific creative processes that are open to iteration and central for the concept of Design Thinking. Weblogs also have the ability to pool information that was so far unshared - another established indicator for successful teamwork, which is the basis of what is often described by creating something larger than the sum of the individual input. We also argued that weblogs support the active solidarization of the individuals into one team, following the similarity-attraction paradigm. Next to the benefit of understanding why and how blogs can support better design thinking teamwork results, one may also retrieve extensive process and decision documentation of rather poorly

investigated design thinking projects. Especially researchers can profit from the possibilities for convenient retrieval of stored expertise. Additionally, the information can support and speed up coaching and learning - new students can in fact build on the research and ideas of others. We thus argue that weblogs - if implemented correctly in a specific context of application - can indeed improve the innovative process of Design Thinking greatly.

REFERENCES

- [1] T. Brown, "Design Thinking," *Harvard business review*, vol. 86, no. 6, pp. 84–92, 141, Jun. 2008. [Online]. Available: <http://www.ncbi.nlm.nih.gov/pubmed/18605031>
- [2] S. Jackson, A. Joshi, and N. L. Erhardt, "Recent research On Team and Organizational Diversion: SWOT Analysis and Implications," *Journal of Management*, vol. 29, no. 6, pp. 801–830, 2003.
- [3] D. Gigone and R. Hastie, "Proper Analysis of the Accuracy of Group Judgments," *Psychological Bulletin*, vol. 121, no. 1, pp. 149–167, 1997.
- [4] J. Lipnack and J. Stamps, *The teamnet factor - Bringing the Power of Boundary Crossing into the Heart of Your Business*. Essex, Great Britain: Oliver Wright Publications, 1993. [Online]. Available: http://www.netage.com/pub/books/download_ttnf.html
- [5] H. W. J. Rittel, "On the planning crisis: systems analysis of the first and second generations," *Bedriftsokonomien*, vol. 8, pp. 390–396, 1972. [Online]. Available: <http://courses.cs.vt.edu/cs4634/reading/Rittel.pdf>
- [6] N. Ross, "The nature and nurture of design ability," *Design Studies*, vol. 11, no. 3, pp. 127–140 (see page 137), 1990.
- [7] B. Lawson, *How Designers Think*, fourth ed. Oxford: Architectural Press, 2006.
- [8] P. G. Rowe, *Design Thinking*. Cambridge, Mass.: The MIT Press, 1987.
- [9] D. Dunne and R. Martin, "Design Thinking and How It Will Change Management Education : An Interview and," *Management Learning*, vol. 5, no. 4, pp. 512–523, 2006.
- [10] H. Plattner, C. Meinel, and U. Weinberg, *Design Thinking*. München: mi-Wirtschaftsbuch, 2009.
- [11] H. W. J. Rittel and M. M. Webber, "Dilemmas in a general theory of planning," *Policy Sciences*, vol. 4, no. 2, pp. 155–169, Jun. 1973. [Online]. Available: <http://www.springerlink.com/index/10.1007/BF01405730>
- [12] G. Hirsch Hadorn, H. Hoffmann-Riem, S. Biber-Klemm, W. Grossenbacher-Mansuy, D. Joye, C. Pohl, U. Wiesmann, and E. Zemp, *Handbook of transdisciplinary research*, G. Hirsch Hadorn, H. Hoffmann-Riem, S. Biber-Klemm, W. Grossenbacher-Mansuy, D. Joye, C. Pohl, U. Wiesmann, and E. Zemp, Eds. Springer, 2008. [Online]. Available: <http://books.google.com/books?id=FzM5FtqBHxoC\&pgis=1>

- [13] S. Song, A. Dong, and A. M. Agogino, "Time Variation of Design "Story Telling" in Engineering Design Teams," in *International Conference on Engineering Design (ICED 03)*. Stockholm, Sweden: ICED, 2003, pp. 1–10.
- [14] G. Stasser, S. I. Vaughan, and D. D. Stewart, "Pooling Unshared Information: The Benefits of Knowing How Access to Information Is Distributed among Group Members," *Organizational Behavior and Human Decision Processes*, vol. 82, no. 1, pp. 102–116, May 2000. [Online]. Available: <http://linkinghub.elsevier.com/retrieve/pii/S0749597800928905>
- [15] C. Noweski, O. Böckmann, and C. Meinel, "The Genesis of a Comprehensive Design Thinking Solution," in *Proceedings of the 17th International Conference on Engineering Design (ICED'09)*, Vol. 7, M. Norell Bergendahl, M. Grimheden, L. Leifer, P. Skogstad, and U. Lindemann, Eds. Stanford, USA: The Design Society, 2009, pp. 217–228. [Online]. Available: http://www.designsociety.org/index.php?menu=40&action=5&search=&_search=&sort_by=publication_year&sort_order=desc&keyword=true&page=31
- [16] D. Rodriguez and D. Solomon, "The Singular Insights of Many Minds," *Innovations*, vol. 2, no. 3, pp. 1–14, 2007.
- [17] J. Broß, K. Richly, P. Schilf, and C. Meinel, *Social Physics of the Blogosphere: Capturing, Analyzing and Presenting Interdependencies of Partial Blogospheres*. New York / Wien: Springer Verlag, 2010, pp. 179–198.
- [18] M. Ojala, "Blogging: For knowledge sharing, management and dissemination," *Business Information Review*, vol. 22, no. 4, pp. 269–276, 2005. [Online]. Available: <http://bir.sagepub.com/cgi/doi/10.1177/0266382105060607>
- [19] J. Broß, H. Sack, and C. Meinel, "Encouraging Participation in Virtual Communities: The IT-summit-blog Case," *Proceedings of IADIS e-Society2007, Lisbon, Portugal, July*, vol. 5, no. 2, pp. 113–129, 2007. [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.2042&rep=rep1&type=pdf>
- [20] H. Kircher, "Web 2.0 - Plattform für Innovation," *it - Information Technology*, vol. 49, no. 1, pp. 63–65, 2007. [Online]. Available: <http://dblp.uni-trier.de/db/journals/it/it49.html#Kircher07>
- [21] J. Broß, A. E. Acar, P. Schilf, and C. Meinel, "Spurring Design Thinking through Educational Weblogging," in *2009 International Conference on Computational Science and Engineering*. Vancouver, Canada: IEEE, 2009, pp. 903–908. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5283071>
- [22] S. Downes, "Educational blogging," *Educause review*, vol. 39, no. October, pp. 14–27, 2004. [Online]. Available: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Educational+Blogging\#0>
- [23] J. B. Williams and J. Jacobs, "Exploring the use of blogs as learning spaces in the higher education sector," *Australasian Journal of Educational Technology*, vol. 20, no. 2, pp. 232–247, 2004. [Online]. Available: <http://www.ascilite.org.au/ajet/ajet20/williams.html>
- [24] A. R. Dennis and J. S. Valacich, "Computer brainstorms: more heads are better than one," *Journal of applied psychology*, vol. 78, no. 4, pp. 531–537, 1993. [Online]. Available: <http://cat.inist.fr/?aModele=afficheN&cpsid=4884366>
- [25] P. B. Paulus and M. T. Dzindolet, "Social influence processes in group brainstorming," *Journal of Personality and Social Psychology*, vol. 64, no. 4, pp. 575–586, 1993. [Online]. Available: <http://psycnet.apa.org/journals/psp/64/4/575>
- [26] T. Kayser, *Team power: How to Unleash the Collaborative Genius of Work Teams*, first edit ed. New York: McGraw-Hill, 1994.
- [27] L. Efimova and S. Fiedler, "Learning webs: Learning in weblog networks," in *Proceedings of the IADIS International Conference Web Based Communities*, no. March 2004, Lisbon, Portugal, 2004, pp. 490–494. [Online]. Available: <http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:LEARNING+WEBS++LEARNING+IN+WEBLOG+NETWORKS\#0>
- [28] M. Razavi and L. Iverson, "A grounded theory of information sharing behavior in a personal learning space," in *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work (CSCW'06)*. New York, New York, USA: ACM Press, 2006, pp. 459–468. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1180875.1180946>
- [29] S. D. Farnham, P. T. Brown, and J. L. Schwartz, "Leveraging social software for social networking and community development at events," in *Proceedings of the fourth international conference on Communities and technologies - C&T '09*. New York, New York, USA: ACM Press, 2009, pp. 235–244. [Online]. Available: <http://portal.acm.org/citation.cfm?doi=1556460.1556495>
- [30] K. Leggett, "The effectiveness of categorical priming in brainstorming," 1997.
- [31] E. Mannix and M. A. Neale, "What Differences Make a Difference ? The Promise and Reality of Diverse Teams in Organizations," *Society*, vol. 6, no. 2, pp. 31–32, 2005.
- [32] HPI, "Innovationslabor: Design Thinking bringt CeBIT-Gäste auf neue IT-Ideen," 2009. [Online]. Available: <http://www.hpi.uni-potsdam.de/presse/mitteilung/beitrag/innovationslabor-design-thinking-bringt-cebit-gaeste-auf-neue-it-ideen.html>
- [33] —, "Hasso-Plattner-Institut: HPI School of Design Thinking," 2010. [Online]. Available: http://www.hpi.uni-potsdam.de/d_school/
- [34] M. Gekeler, "Youre a buddy, shes a buddy everybody is a bring.BUDDY," 2009. [Online]. Available: <https://d-school-blog.hpi-web.de/?p=929>
- [35] N. Birger, ""Bring Buddy" soll Verkehrschaos austricksen," 2010. [Online]. Available: <http://www.welt.de/wirtschaft/article10390463/Bring-Buddy-soll-Verkehrschaos-austricksen.html>
- [36] WDR-Fernsehen, "Dittsche - Der Bringbuddy," 2010.

- [37] W. Davies, "You don't know me, but... social capital & social software," *ISociety*, vol. 3, no. May, pp. 1–64, 2003. [Online]. Available: <http://blogoehlert.typepad.com/eclippings/files/1843730103.pdf>
- [38] B. Nardi, S. Whittaker, and H. Schwarz, "NetWORKers and their activity in intensional networks," *Computer Supported Cooperative Work (CSCW)*, vol. 11, no. 1, pp. 205–242, 2002. [Online]. Available: <http://www.springerlink.com/index/BJ722AJU32Q7D4FL.pdf>
- [39] E. C. U. Pooley, Julie Ann (School Of Psychology, L. Cohen, and L. Pike, "Can sense of community inform social capital?" *The Social Science Journal*, vol. 42, no. 1, pp. 71–79, 2005. [Online]. Available: http://www.sciencedirect.com/science?_ob=ArticleURL&_udi=B6W64-4F29J2M-5&_user=1584062&_coverDate=01/01/2005&_rdoc=1&_fmt=high&_orig=search&_sort=d&_docanchor=&view=c&_acct=C000053886&_version=1&_urlVersion=0&_userid=1584062&md5=c6baa377372d2ef1bba1d546ad7949a6
- [40] M. L. Forret and T. W. Dougherty, "Correlates of Networking Behavior for Managerial and Professional Employees," *Group & Organization Management*, vol. 26, no. 3, pp. 283–311, Sep. 2001. [Online]. Available: <http://gom.sagepub.com/cgi/content/abstract/26/3/283>

Promoting organisational emergence in business social networks

Matei Dobrescu

IT General Directorate
Insurance Supervisory Commission
Bucharest, Romania
mdobrescu@csa-isc.ro

Abstract— In this paper complex adaptive systems theory (CAS) and social autopoiesis have been interpreted with the aim to identify factors realising emergent properties in organisations structured as social networks. Understanding the complex dynamics of such communities requires a view of their infrastructure as a network of interacting agents involving both goals and constraints. We analyze the network structure, showing that it defines a complex weighted network with scaling laws at different levels. We also present a simple model of network growth involving non-local rules.

Keywords - Social Networks; Self – organization; Complex Adaptive Systems; Organisational emergence; Social autopoiesis

I. INTRODUCTION

Looking back there have been a few distinctive technological innovations that have radically changed the way society operates and is able to interact. Johannes Gutenberg's printing press is the earliest example. For the first time knowledge could be shared to a wider population and what this did was take away the control of knowledge from the nobility and transfer it to the general population, this was the first step towards the democracy of knowledge. Radio and television and the Internet are more recent examples but the most exciting step, social networking has just arrived and will once again have a major impact on all elements of society.

The Internet has been around for almost 20 years and has made drastic changes in the way we run our lives and the manner in which we conduct our business. The internet has provided us with a platform to exchange information and create knowledge in exponential quantities. We are however only now beginning to truly collaborate globally in how we exchange and create knowledge.

One place that we are beginning to see this kind of global collaboration and knowledge creation is in social networks. It took radio 28 years to reach a market audience 50 million, 13 years for TV, 4 years for the internet and only 2 years for FaceBook.

More and more of our personal data are making it onto the web every day. From applications as pedestrian as word processing to social networking tools such as Loopt, which allow one to share their GPS location with friends, the web is supplanting the classic Personal Computing paradigm. The

smart phone is accelerating this trend. Users expect to be able to view data produced on the desktop while on the go.

Centralized application services have many positive properties. They are easy to use. They make it easy to share. It is fun and easy to discover new friends, to discover who your friends have as friends or to reconnect with old friends. Users do not need to worry about software upgrades as the application provider automatically updates the software as needed. Third party application developers have, through independent development, made these platforms more useful and fun than ever before.

All of this freedom does come at a cost however. The risks created by centralized service providers is worthy of concern, and one can consider that in less than five years a network like Facebook is worth more as a direct mail marketing service than it is as a social networking application. Even worse seems to be the freedom in social networks communication for businesses. With a digitally connected social world in which the line between personal and corporate lives is increasingly blurred, potential risks to businesses also rise. It's clear that businesses need a proactive— and powerfully persuasive—communications plan to educate their user community about social media risks, personal and company impacts, and expected behaviors.

On the other hand, despite the risks, many companies are ill-prepared. To safeguard critical data, mitigate data leakage, and control intellectual property, one must adopt a strategy that leverages the experience and leadership of the business and technology sides of the companies. As an alternate solution, we propose to use in social network management a distributed platform that retains the core functionalities of a centralized service with the additional advantage of returning ownership of the data to the user. The existence of a distributed solution offers consumer choice and puts pressure on centralized services to treat our data with the care and discretion we desire.

II. THE POLICIES AND PROCESSES FOR SUCCESSFUL SOCIAL NETWORKS

As with any policy implementation, the first step concerning social media is to form a business strategy that includes a long-term adoption plan for policies, procedures, and solutions.

It is essential that the business classify data so that employees understand precisely what is—and is not—sensitive information. This process also should define who is authorized to access and share corporate content, and it should lay out procedures that delineate how employees may use sensitive data. As part of data classification, the business should also establish a data-retention policy for information created on social media.

Policy also must clearly specify who is responsible for particular types of communications; these operational roles typically fall within the marketing and customer service departments. The company also should establish management oversight for social media—both a chief strategist and a community manager, for instance.

When developing roles and policies, the business should include a strategy for employee separation to maintain ownership of intellectual property and social identities.

Establishing these policies is only the beginning, however. The real work lies in behavioral changes of employees. Businesses must educate employees on the need to protect intellectual property and sensitive information, and they should fully detail the consequences of noncompliance for both the company and the individual.

Typically, technology incubated by Computer Science professionals in universities and companies eventually make its way to consumers. Distributed systems have not made this leap. Consumers have the same need to share media with friends over the Internet. We envision that personal servers of tomorrow may become as prevalent as today's personal computers. Obviously we still have a long way to go before today's social applications. The key is to create open high-level distributed programming interfaces and frameworks that enable independent software vendors to create distributed applications that run across these servers.

Privacy, the key factor in our new design, must make possible a new class of viral applications and preserve and even enhance the ability of advertisers to make a profit. Without privacy, an entire class of financial and medical applications will not be accepted. In fact, privacy is also useful for applications involving interpersonal relationships, a particularly viral category. While it is generally accepted that the younger generation has less qualms over making personal information public, few would be willing to make public their negative feelings about other individuals.

Concluding that in turbulent business environments organisations need to react quickly and creatively to make the most of new opportunities and business models, in this paper we consider as a possible solution to reach these new imperatives which require organisations to become more flexible to handle change, the model of complex self-organizing systems, where of key importance in responding successfully to change is the concept of emergence. Complexity science is a way of addressing and improving such capabilities in organisations, as it is concerned with the role of chance, emergence and contingency in the face of frequent and continuous change [1].

In our work factors facilitating organisational emergence have been identified by interpreting complex adaptive systems (CAS) and social autopoiesis theories with the aim

of identifying mechanisms or strategies that raise the emergent properties of social business enterprises [2]. Social autopoiesis was chosen as it focuses on social elements of emergence, such as communication, collaboration, morale, trust, etc., whereas CAS theory concentrates more on adaptive mechanisms that make a CAS produce emergent order, such as inter-relations, interconnectivity, edge of chaos, feedback, etc. Based on this a framework has been derived that summarises the so-called factors that facilitate organisational emergence. The framework classifies factors as tangible and intangible, and it differentiates between dynamics, enabling infrastructure and controls, amongst emergence factors. By enforcing factors facilitating emergence and avoiding factors prohibiting emergence, it is argued that organisational emergence will be leveraged leaving space to project teams to innovate and continuously evolve appropriate solutions in order to adapt to an ever-changing business environment.

III. CHARACTERIZATION OF THE SOCIAL NETWORKS AS COMPLEX ADAPTIVE SYSTEMS

Self-organization has been subject of discussions concerning the question of the interrelationship between a system and its environment in various disciplines, apart from DAI (Distributed Artificial Intelligence). The different theoretical approaches have in common that they call any kind of system self-organizing if it is able to determine its internal structure by itself as the environment changes. The boundaries of a self-organizing system and its structure (i.e. the relation between its elements) are not determined by environmental factors. Rather, these systems generate, change and adapt their internal organization within their own logic in a dynamic process to cope with environmental changes. As a result of more recent social theories, the notion of self-organization has become a primitive in sociology when it comes to describe social entities (groups, networks, organizations). In particular in MAS (Multiagent Systems) literature the concept of self-organising MASs has been partially considered by researchers interested in designing the best match among task, environment, structure and performance. A prevalent opinion is that sociological theory can help overcoming difficulties in modeling MAS. In this spirit it is to mention a new sociological concept to the study of self-organization in MAS, the habitus-field theory of Pierre Bourdieu which describe organizations as self-organizing social entities ("autonomous fields") [3].

Most of the work on complexity and the development of complexity theories have been undertaken in the context of the natural sciences and there has been relatively little work on developing or applying such theories in the social sciences. A thorough review of complexity and social autopoiesis literatures is undertaken in this section, based on the work of Alaa [4], with special focus on management-related contributions to extract mechanisms or groupings of factors that are argued will facilitate emergence in social and management contexts. The analysis resulted in a classification into several groupings; dynamics (social construction factors/intangible dynamics and adaptive

factors/tangible dynamics), enabling infrastructure (tangible & intangible), and control factors (tangible & intangible). Dynamics are factors that realise emergent properties, the enabling infrastructure include elements that enable the dynamics to become effective, whereas controlling factors attempt to ensure balance of dynamics to prevent descent into chaos.

A. Social Construction Factors/Intangible Dynamics

The social drivers and stimulators that have been suggested as important in facilitating emergent social behaviour are presented as follows:

- The development of autopoietic society requires communication, meaning and consciousness that form an essential driver of emergent behaviour.
- Facilitation of interaction in the development of social organisations put co-operative interaction and relationships at the centre of organisational emergence, which can be achieved through participation, collaboration and team working.
- Local interactions are responsible for new order creation and emergence of global structures.
- The quality of interactions between human agents is a function of the diversity, density, and intensity of those relations. These may be formal or informal, designed or un-designed, implicit or explicit
- Individual motives or intentions and individual emotions and morale act as driving forces for social autopoietic systems influenced by interests, social context and forms of co-operation and collective behaviour towards achieving a specific goal.

Thus, the important social construction factors are communication, collaboration, interaction, trust and morale. These appear to be the important elements of complex social systems as they are responsible for social interactions and stimulation of creative thinking that will lead to human empowerment and leveraging self-organisation.

B. Adaptive Factors/Tangible Dynamics

The dynamic of an evolving social entity is determined by inter-component relationships that outline its form and internal arrangements. Adaptive factors are required to improve the ability of the social system to re-arrange, reform its structure and quickly respond to change; they include the following elements:

- In a social context each individual belongs to many groups and different contexts and the contribution depends partially on the other individuals within that group and the way they interact.
-
- Propagation of influence through social system depends on the degree of connectivity, interdependence and strength of coupling.
- In human systems, connectivity between individuals or groups is not a constant or uniform relationship, but varies over time.

- Complexity thinking is about wholes and complex inter-relationships.
- Difficulties created by the unpredictability of complex human processes and interdependencies are problematic, therefore short-term orientation and simple solutions (simplicity) are likely to result in better outcomes.
- Conditions for experimentation and exploration of possibilities implies small-scale orientation in order to quickly try out various options and get quick feedback without requiring large scale resources and time.

Thus, the adaptive factors reflect the degree of interdependence, connectivity, structural coupling and quick re-formation of internal arrangements. These elements help facilitate fast response and quick, internal adaptation and re-formation of system components.

C. Enabling Infrastructure

Aspects of an enabling infrastructure that facilitates emergence in social contexts include:

- Hierarchy and structure are pre-conditions that enable or inhibit the emergence of new behaviours and working ways.
- Action of organisation members is shaped to a high degree by the existence of specific organisational form and structures.
- Conditions that facilitate the day-to-day management of an organisation, for example management style, are necessary for learning and emergence to occur.
- Analysis of the influence of external factors like power, money and control regulations like contracts and conventions act as constraints that limit social dynamics in complex situations.

D. Control Factors

Complexity theory in social contexts is designed to enable creativity, spontaneity and emergence but it also requires some kind of moderating or control mechanisms, which seeks to balance excessive change with stability, possibilities with constraints, innovation with tradition, etc.

- Change and stability are balanced and the edge of chaos is a critical point of the system, where a small change can either push the system into chaotic behaviour or tip the system back into a stable state.
- Edge of chaos is controlled by equilibrium models which attempt to bound a system to ensure that the system is always pushed back to stable conditions.
- The mechanisms by which complex systems maintain control and achieve certain goals is by feedback, learning and frequent small adjustments to counteract any excessive tendencies to change.
- Continuous reflection, learning and circular causality mutually reinforce social relationships and interactions.
- Simple high-level rules are a way to achieve a balance between dictation and freedom enabling

team members to interact with each other guided by these rules

Based on the above analysis we identify the first grouping of factors facilitating emergence, i.e. dynamics that include those factors that operationalise the emergent behaviour. The factors of a complex social system are also classified into intangibles and tangibles. Intangibles represent the social factors that uniquely characterise social human systems, as opposed to natural systems, whereas tangibles represent the mechanistic/adaptive factors, those elements responsible for the internal connectivity of system components.

The second grouping is the enabling infrastructure that enables or allows the social and adaptive elements to either be effective or inhibited. This includes organisational structure, hierarchies, management style, work culture, leadership, etc. These elements can also be tangible, such as structures, hierarchies and external factors or intangible, such as culture, management style and leadership. The third grouping is control, as in order to facilitate emergent behaviour without complete chaos or anarchy, controls need to be in place and maintained, but they need not to be too restrictive. The different groups and elements of each category are illustrated in Fig. 1 that collates the various factors. It is argued that this forms a useful framework for identifying and understanding factors that facilitate organisational emergence.

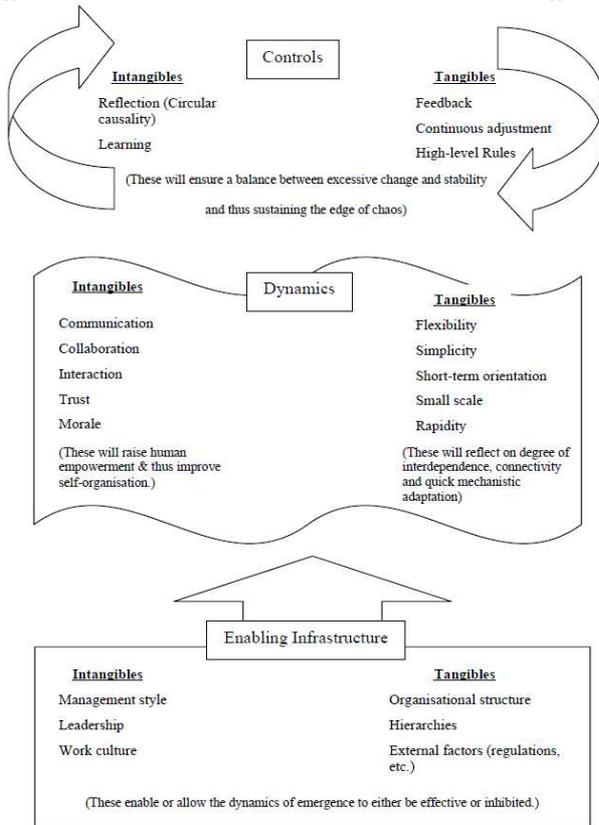


Figure 1. Framework of Factors Facilitating Organisational Emergence

IV. AN INFRASTRUCTURE MODEL FOR SOCIAL NETWORKS

Social network analysis represents agent relationships with nodes and links. Every node i represents an actor i within the network and links (i, j) denote social ties between agents i and j . More representative models of social networks decorate each link (i, j) with the strength of the social tie or the amount of information flowing through it, hereafter called link weight $w_{i,j}$. The statistical analysis of link weights $w_{i,j}$ between pairs of vertices in the social network indicates an heterogeneous pattern of interactions, typically following a power law: $P(w_{i,j}) \sim w_{i,j}^{-\lambda}$. In addition, the heterogeneous distribution of link weights might be related to the hierarchical organization of the social network.

We have chosen as specific example for which it is possible to reconstruct the social network, the social network of open source software communities (OSS). This system define a network of interacting agents with very similar features in common, reflecting the presence of limitations in the information shared by agents. It has been argued that decentralization leads to a distinctive organization that solves the communication bottleneck associated to large software projects [5]. The amount of submitted e-mails from one programmer to other members is a good indicator of his social position in the software community. However, not every e-mail message has the same influence in the process of software development. In order to reduce the amount of noise, here we will consider only e-mail traffic associated to bug-fixes and bug reporting. The rest of e-mails are discarded from any further consideration. From this subset of e-mails we can reconstruct the social network of the software community as shown in [6].

Nodes and links (i, j) of the OSS social network represent members and e-mail communication from i to j , respectively. At any time, a new software bug is discovered by the member i who sends a notification e-mail. Then, other expert members investigate the origin of the bug and eventually reply with the solution. Typically, several messages are required to solve the problem. Here, we define $E_{i,j}(t)=1$ if developer i replies to developer j at time t , or $E_{i,j}(t)=0$ otherwise. We also define link weight $w_{i,j}$ as the amount of e-mail traffic flowing from member i to member j ,

$$w_{i,j} = \sum_{t=0}^T D_{i,j}(t) \text{ where } T \text{ is the timespan of software development.}$$

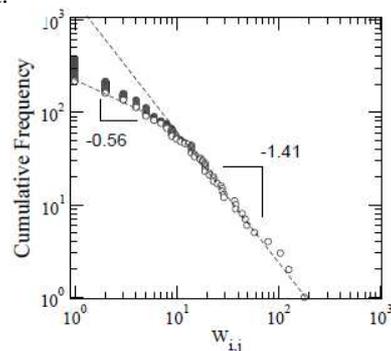


Figure 2. Heterogeneous interaction in small software communities

In fig.2 we put an emphasis in the link weight distributions $P(w_{i,j})$. Here $P_{>}(w_{i,j})$ is defined as the probability of having a link with weight $w_{i,j}$. In order to reduce the noise in the statistical data, we make use of the cumulative distribution $P_{>}(w_{i,j})$, defined as

$$P_{>}(w_{i,j}) \equiv \int_{w_{i,j}}^{\infty} P(\omega) d\omega .$$

For the standard case where a scaling behaviour $P(w_{i,j}) \sim w_{i,j}^{-\lambda}$ is observed, we have $P_{>}(w_{i,j}) \sim w_{i,j}^{-\lambda+1}$.

There is a characteristic pattern of asymmetric interaction, where a few strong units dominate the activity of the whole OSS. Interestingly, the distribution of link weights in large software communities also follows a power-law; with an exponent consistent with the observed in the small software communities. Most real networks typically contain parts in which the nodes (units) are more highly connected to each other than to the rest of the network. The sets of such nodes are usually called clusters, communities, cohesive groups, or modules having no widely accepted, unique definition.

In general, each node i of a network can be characterised by a membership number m_i , which is the number of communities the node belongs to. In turn, any two communities α and β can share $s_{\alpha,\beta}^{ov}$ nodes, which we define as the overlap size between these communities. Naturally, the communities also constitute a network with the overlaps being their links. The number of such links of community α can be called as its community degree, d_{α}^{com} . Finally, the size of any community α can most naturally be defined as the number of its nodes. To characterise the community structure of a large network we introduce the distributions of these four basic quantities. In particular, we will focus on their cumulative distribution functions denoted by $P(s^{com})$, $P(d^{com})$, $P(s^{ov})$, and $P(m)$, respectively.

The basic observation on which our community definition relies is that a typical community consists of several complete (fully connected) subgraphs that tend to share many of their nodes. Thus, we define a community, or more precisely, a k -clique-community as a union of all k -cliques (complete subgraphs of size k) that can be reached from each other through a series of adjacent k -cliques (where adjacency means sharing $k-1$ nodes). This definition is aimed at representing the fact that it is an essential feature of a community that its members can be reached through well connected subsets of nodes. There are other parts of the whole network that are not reachable from a particular k -clique, but they potentially contain further k -clique-communities. In turn, a single node can belong to several communities. All these can be explored systematically and can result in a large number of overlapping communities. Notice that in most cases relaxing this definition (e.g., by allowing incomplete k -cliques) is practically equivalent to lowering the value of k . In the same time any k -clique (complete subgraph of size k) can be reached only from the k -cliques of the same community through a series of

adjacent k -cliques (two k -cliques are adjacent if they share $k-1$ nodes) [7].

The algorithm for numerical determination of the full set of k -clique-communities is based on first locating all cliques (maximal complete subgraphs) of the network and then identifying the communities by carrying out a standard component analysis of the clique-clique overlap matrix [8]. We use our method for binary networks (i.e., with undirected and unweighted links). An arbitrary network can always be transformed into a binary one by ignoring any directionality in the links and keeping only those that are stronger than a threshold weight w^* . Changing the threshold is like changing the resolution with which the community structure is investigated: by increasing w^* the communities start to shrink and fall apart. A very similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated, but at the same time, also more cohesive.

The extent to which different communities overlap is also a relevant property of a network. Although the range of overlap sizes is limited, the behaviour of the cumulative overlap size distribution $P(s^{ov})$ is close to a power law for each network, with a rather large exponent. This remark led us to consider that the most suitable topology of a social network is that of a scale free network [9].

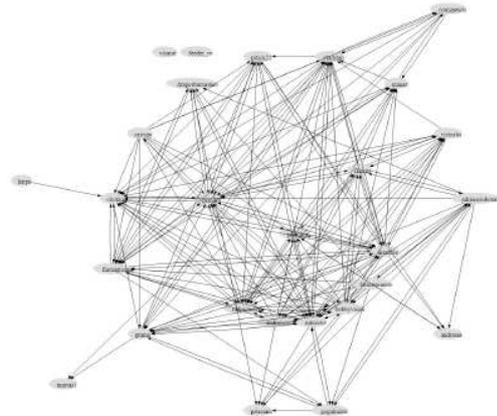


Figure 3. The network of the most influential 27 blogger profiles on Twitter

In order to establish whether social networks are indeed scalefree, we determined the degree distribution $P(k)$, which is the probability of finding a node with a degree k in the Romanian Blogosphere (the interconnected network of Romanian bloggers). The obtained distribution is indeed scale-free and satisfies the power law with the exponential: $\lambda=2.65$ which satisfies the condition to be between 2 and 3 for a scale-free topology. As expected, the most influential persons in the Romanian Blogosphere will also have accounts on a large social network as Twitter and will keep their superiority there also. From the first 100 blogs, in July 2008, already 27 were also on Twitter. Figure 3 shows the interconnection between these most influential blogger profiles which are also interconnected on Twitter. The scale-free topology following the preferential attachment law is easy to observe.

To determine the connectivity degree of such a network, we have made simulations using Ns2 simulator [910] and the Nam animation tool [11]. Fig.4 illustrates the topology of a free scale network with 128 nodes that started from an initial core of 4 nodes; in the connection of other nodes we have applied the law of the preferential attachment. In fig. 5 is represented the distribution of the connectivity degree from the most connected node till the less connected one, for 4 scale free networks with nodes from 100 to 100000, the similarity being evident.

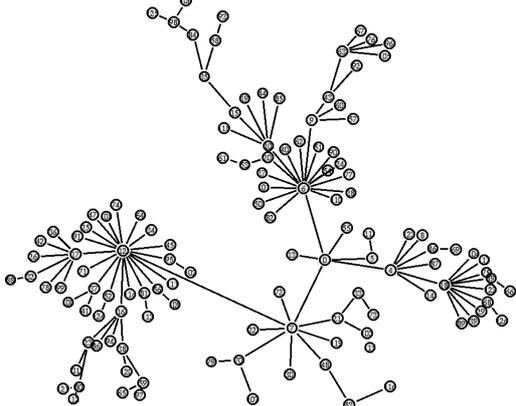


Figure 4. A scale free network with 128 nodes having 5 hubs

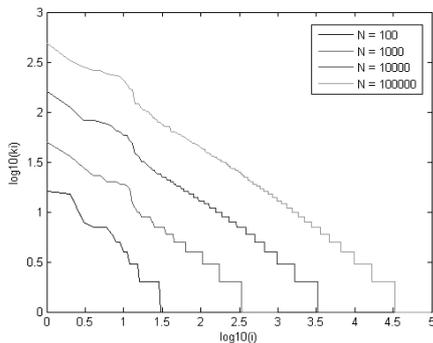


Figure 5. The evolution of the node connectivity for 4 free-scale networks

The specific scaling of the community degree distribution is a novel signature of the hierarchical nature of the systems we study. We find that if we consider the network of communities instead of the nodes themselves, we still observe a degree distribution with a fat tail, but a characteristic scale appears, below which the distribution is exponential [12].

CONCLUSION

In this paper complex adaptive systems theory (CAS) and social autopoiesis have been interpreted with the aim to identify factors realising emergent properties in organisations defined as social networks. Social construction elements, such as communication, collaboration, interaction, trust, etc. are argued to be critical drivers of human empowerment and thus self-organisation, whereas mechanistic, adaptive

dynamics like flexibility, short-term orientation, small scale approaches, simplicity and rapidity will ensure fast response and quick adaptation to the problem situation. However, emergence cannot be fully realised without the necessary enabling infrastructure that will allow the dynamics of emergence to become effective, e.g. management style, work culture, organisational structure etc. The elements or factors in each category have been identified and related in a framework, to help understand and analyse the phenomenon of emergence in social organisations.

This framework can be seen as a significant improvement on generic complexity principles suggested in literature, such as diversity, large number of agents, interactions, edge of chaos, etc. that refer to emergence characteristics but without providing a clue on how to realise these concepts in action.

Especially, it is important to notice that the framework represents a holistic approach where the various identified factors are intertwined and some of them may produce counteracting effect. Future research will focus on further validation of the framework through other empirical applications. Especially of interest is to test if the framework does help better understand and manage the emergence phenomenon and put forth intentionally factors that raise the emergence of new work arrangements. Generality and completeness of the framework are also important to test in future work.

REFERENCES

- [1] E. Mittleton-Kelly, "Ten Principles of Complexity and Enabling Infrastructures", in Mittleton-Kelly (ed.), *Complex Systems and Evolutionary Perspectives on Organisations*, Elsevier, 2003.
- [2] C. Hauert, "Cooperation, collectives formation and specialization", *Advances in Complex Systems*, Vol. 9, No. 4, 2006, pp. 315-335
- [3] P. Bourdieu, *Pascalian Meditations*, Cambridge, Polity Press, 2003
- [4] G. Alaa, "Derivation of Factors Facilitating Organizational Emergence Based On Complex Adaptive Systems and Social Autopoiesis Theory", *E:CO*, Vol. 11. No. 1. 2009, pp. 19-34
- [5] S. Valverde and R. V. Sole, "Self-organization versus hierarchy in open-source social networks", *Phys. Rev. E*, Volume 76. Issue 4, 2007, pp. 44-50
- [6] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The Architecture of Complex Weighted Networks", *PNAS*, vol. 101, no. 11., 2000, pp.3747-3752.
- [7] G. Palla, I. Derényi, I. Farkas and T. Vicsek, "Uncovering the overlapping community structure of complex networks in nature and society", *Nature*, 435, 2005, pp. 814-818
- [8] M. E. J. Newman, "Detecting community structure in networks", *Eur. Phys. J. B*, 38, 2004, pp. 321-330
- [9] R. Dobrescu, D. Hossu, and R. Ulrich. "Self-similarity Tests for Internet Traffic, *Control Engineering and Applied Informatics*, Vol. 11, No. 4, p.11-17, dec. 2009,
- [10] K. Fall and K. Varadhan – *The ns Manual, The VINT Project: A Collaboration between researchers at UC Berkeley, LBL, USC/ISI, and Xerox PARC*, 2008.
- [11] D. Estrin, M. Handley, J. Heidemann, S. McCanne, Y. Xu and H. Yu, "Network Visualization with Nam, the VINT Network Animator", *Computer*, 11, 2000, pp. 63-68
- [12] C. Song, S. Havlin, and H. A. Makse, "Self-similarity of complex networks", *Nature* 433, 2005, pp. 392-39

Mitigating Risk in Web-Based Social Network Service Selection: Follow the Leader

Jebrin Al-Sharawneh, Mary-Anne Williams, Xun Wang and David Goldbaum*

Quantum Computation and Intelligent Systems (QCIS)

Faculty of Engineering and Information Technology, *School of Finance and Economics

University of Technology, Sydney, Australia

e-mail: jebrin@it.uts.edu.au, Mary-Anne.Williams@uts.edu.au, xuwang@it.uts.edu.au, David.Goldbaum@uts.edu.au

Abstract— In the Service Web, a huge number of Web services compete to offer similar functionalities from distributed locations. Since no Web service is risk free, this paper aims to mitigate the risk in service selection using “Follow the Leader” principle as a new approach for risk-reducing strategy. First, we define the user credibility model based on the “Follow the Leader” principle in web-based social networks. Next we show how to evaluate the Web service credibility based on its trustworthiness and expertise. Finally, we present a dynamic selection model to select the best service with the perceived performance risk and customer risk-attitude considerations. To demonstrate the feasibility and effectiveness of the new “Follow the Leader” driven approach to alleviate the risk in service selection, we used a Social Network Analysis Studio (SNAS) to verify the validity of the proposed model. The empirical results incorporated in this paper, demonstrate that our approach is a significantly innovative approach as risk-reducing strategy in service selection.

Keywords - Web service selection, credibility assessment, risk, web-based social networks, Follow the Leader.

I. INTRODUCTION

In the Service Web, Web services and Web-Based Social Networks will emerge to create an environment where users and applications can search and compose services in an automatic and seamless manner. The Service Web is expected to be a place where a huge number of Web services will compete to offer a wide range of similar functionalities. It is expected that Web services will fully leverage the Semantic Web to outsource part of their functionalities to other Web services [1]. In this case, some services may not have interacted before, while others may act maliciously to be selected. A key requirement is to provide trust mechanisms for quality access and retrieval of relevant Web services with perceived risk considerations.

In the Web service selection, reputation assessment mechanisms are used to establish trust between Web services. The notions of “trust” and “reputation” are both used to evaluate an entity’s trustworthiness [2]. Recent research [3] shows that a good Web service reputation positively affects the consumer’s trust and negatively affects the consumer’s perceived risk. For example, consumers are hesitant to transact with a service provider who has a history of failing to honor its obligations, whereas it is relatively

less risky to transact with a vendor who has a history of honoring its obligations.

Web service selection is a complex process where a service that best satisfies user preferences is selected from a set of candidate services based on user requirements [4] As per the selection criteria, various non-functional properties such as quality of service (QoS), can be used and expressed as user preferences. QoS such as response time, throughput, availability, reliability and privacy are difficult for the user to determine and control. Users are usually not willing to spend time describing their detailed preferences to the system. They are even less inclined to assign weights to them, especially if they do not have a clear understanding of the effects and results of this input. Moreover, users may not even be aware of their explicit preferences. Hence, risk-averse users who want to use Web services often seek help from their friends, peers, experts and business partners who may have relevant expertise or experiences.

In this paper, we propose a service selection approach based on a credibility framework that models user and Web service credibility with Web service perceived risk and user risk attitude. Our work is the first that uses a formal “Follow the Leader” model [5] based on web-based social networks and service credibility to mitigate risk in service selection using the most trustworthy and experienced users in the social network.

In order to simplify the paper, we will refer to customers / users as human users, and a Web service as an atomic service such as a home loan or a home insurance service. The proposed approach can be used as a module of Web services personalized recommender system where user behavior can be captured from his/her interactions in WBSN.

II. MOTIVATIONS AND CONTRIBUTATIONS

Decision making in risky complex situations has always been a very difficult task. Traditional decision models for Web service selection based on utility only are no longer adequate; service selection is more complicated with traditional approaches because the consumers may not even know with whom they are interacting.

To illustrate the challenges involved in Web service selection we provide the following example, which illustrates the key difficulties and at the same time motivates our approach.

Motivating Example:

Bob just moved to the USA. By nature he is a risk-averse person. He is seeking an insurance company to insure his home. Bob lives in the same area as his friend Adam who has already taken out home insurance. This is Bob's first house, and he does not want to spend too much time on analyzing insurance features he would rather have the same insurance as his friend Adam. What if Bob did not know Adam? Can he get reasonable advice from somebody who lives in his area? If not, then he would have to embark on tedious and time consuming process of differentiating between the vast number of home insurance services in which all of them may match his request from a functionality perspective, but vary in their non-functional properties.

Using trust in social networks provides a promising approach to make recommendations to other users based on trust propagation in finding a friend or a friend of a friend with similar interests. However, even when the user relies on a trustworthy friend there is still an amount of perceived risk to be considered in adopting the Web service recommended. The quality of the selected Web service can be improved further by assessing its credibility by incorporating its trustworthiness and expertise at the same time. Our key contribution in this paper is threefold:

1. A user model with risk-attitude based on user credibility that captures trust relationships between users.
2. A Web service credibility metrics that incorporate trustworthiness, expertise and perceived risk.
3. A Web service selection approach based on the service credibility and Follow the Leader to mitigate the performance risk in service selection.

The rest of this paper is organized as follows: Section III presents a review of some related works. In Section IV we propose a credibility based framework, next in Section V we model perceived risk and risk attitude in Web service selection, followed by simulations. Finally, we conclude by summarizing our findings and future plans for further work.

III. RELATED WORK

In the following section, we present the synergies which are used in our framework.

A. Web-Based Social Networks and Trust

Web-based social networks (WBSNs) are online communities "people, organizations or other social entities" [6] connected by a set of social relationships, such as friendship, co-working or information exchange in varied contexts e.g., entertainment, religion, dating, or business.

Over the last few years, interest in social networking websites such as MySpace, Twitter and Facebook have increased considerably [7]. Hundreds of millions of people are members of social networks online and many of those networks contain trust data [8]. With access to this information, trust has the potential to improve the way recommendations are made and services are selected.

In WBSNs, the trust inference mechanism is becoming a critical issue when participants want to establish a new trust relation or measure trust values between connected users [9]. The idea is to search for trustworthy users by exploiting trust propagation [10] over the trust network.

B. Trust and Risk in Service Selection

Trust and risk are two tools for making decisions in uncertain environments [11]. In such environments, where the service consumer often has insufficient information about the service provider and the offered services, this forces the consumer to accept the risk of prior performance [12], i.e., to pay for services before receiving them, which can leave her in a vulnerable position. Trust comes into play as a solution for the specific problems of risk. Trust becomes the crucial strategy for dealing with an uncertain and uncontrollable future. So, trust is particularly relevant in conditions of ignorance or uncertainty with respect to the unknown actions of others.

There are only a few computational trust models that explicitly take risk into account. Studies that combine risk and trust include [13] and [11]. In PET, Liang and Shi [13] their conclusion highlights that risk is important in designing a personalized trust system.

Trust can be described as a positive state of mind caused by the perception that the risk resulting from collaborating with the trusted party is acceptable [14]. Trust systems enable parties to determine the trustworthiness of participating parties. Trust is relevant in situations where one must enter into risks but has incomplete control over the outcome, hence any act of trusting implies some bet and some risk [15]. A recent study [3] concludes that as trust increases, consumers are likely to perceive less risk than if trust were absent; i.e., the consumer's trust negatively affects the consumer's perceived risk of a Web service transaction.

C. Follow the Leader

As pointed out by social psychology theory [9], the role of a person in a specific domain has significant influences on trust evaluation if the person recommends a person or an object. Follow the leader in dynamic social networks [5], is a formal probabilistic model of opinion formation with dynamic confidence in agent-mediated social networks where the profiling of agents as leaders or followers is possible. An opinion leader is specified as a highly self-confident agent with strong opinions. According to [5], in a social network, a member is either a leader or a follower who adopted another leader's opinion to use a Web service. Subsequently this member adopts whatever her best friend adopted, otherwise the member has no active friends and consequently it acts as an independent user.

Ramirez-Cano and Pitt [16], define the relationship between two agents as a confidence function, such that: "an agent (i) increases its confidence in another agent (j) based on how well (j's) opinion meets the criteria specified in i's

mind-set. A mind-set represents the set of beliefs, attitudes, assumptions and tendencies that predetermine the way an agent evaluates a received opinion”.

IV. CREDIBILITY BASED FRAMEWORK

A. Web Based Social Network (WBSN) Interaction Model

In a WBSN, as shown in Fig. 1, let a set of users $U = \{u_1, \dots, u_N\}$ interacting in a set of contexts or domains $D = \{d_1, \dots, d_L\}$, such as categories in EPINIONS.com. In each domain there is a set of Web services (K), such that: $S = \{S_1, \dots, S_K\}$, where $S \in D$.

Each user ($u \in U$) rates a set of Web services M denoted by: $R_u^S = \{R_u^1, \dots, R_u^i, \dots, R_u^M\}$, where $M \leq K$, and (R_u^i) is the rating value of user u for Web service S_i . The rating value can be any real number, but most often ratings are integers, e.g., in the range [1, 5].

In a trust-aware system, there is also a trust network amongst users. We define (T_u^v) to be the direct trust between user u and user v , trust value is a real number in the range [0, 1]: 0 means no trust and 1 mean full trust between users.

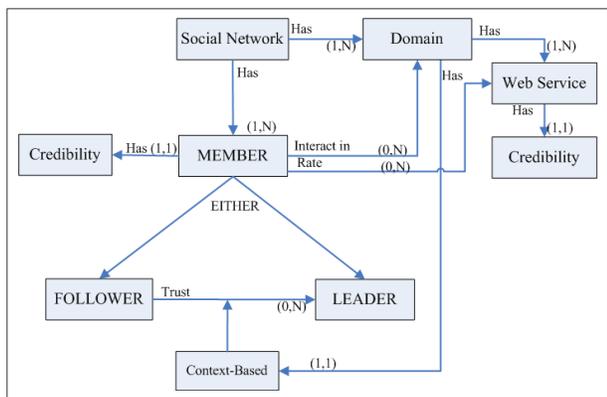


Figure 1. WBSN user interaction model

B. User Credibility Based Clustering - Follow the Leader

The “Follow the Leader” model [5], provides us with insights to identify users based on their roles in the WBSN i.e., either leaders or followers. Enriching the “Follow the Leader” model with trust, gives us the potential to analyze WBSN based on user’s credibility. Fig. 1 shows the basis of our approach. User credibility measure reflects their trustworthiness and expertise and provides us with the means to identify users’ roles in a specific context. Some users can be classified as leaders others can be classified as followers according to their credibility level.

User credibility is a synonym of believability [17]. Credibility of an agent can be measured by its trustworthiness, expertise, and dynamism [18]. The majority of researchers identify two key components of credibility: trustworthiness and expertise. In our previous work [19, 20]

we derived a formula to express user credibility in a WBSN. This formula is expressed as:

$$Cr(u) = \alpha * Cr(R_u) + \beta1 * Cr(T_D^u) + \gamma * Cr(T_I^u) \quad (1)$$

User credibility components consist of: (1) $Cr(R_u)$ refer to user credibility expertise from user ratings component, (2) $Cr(T_D^u)$ refer to user credibility trustworthiness from direct followers trust and (3) $Cr(T_I^u)$ refer to user credibility trustworthiness from indirect followers trust, where $\alpha + \beta1 + \gamma = 1$, and $\alpha, \beta1, \gamma$ are system tuning parameters representing the importance of each credibility component. In our experiments, we use the values (5/9,3/9,1/9) respectively.

Credibility of Web service is a crucial part in service selection. In the following section we define Web service credibility and show how to compute it in a dynamic environment.

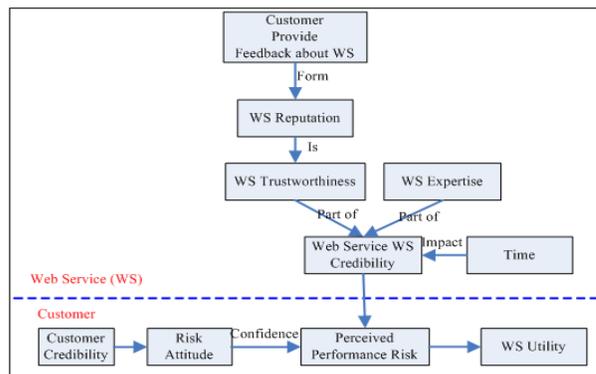


Figure 2. Web service Credibility Model

C. Web service Credibility Model

We define a credible Web service as a service that performed consistently, accurately, and has proven to be dependable over a period of time (t). Credibility of a Web service $Cr_t(S)$ as shown in Fig. 2 can be measured by its trustworthiness $Cr_t(T_S)$, expertise $Cr_t(E_S)$ and dynamism [18]; so we address these components as follows:

1) Web service credibility from Trustworthiness

Trustworthiness is the property of an entity of being “able to be trusted”, while trusting is “to have belief or confidence in the honesty, goodness, skill or safety of a person, organization or thing” [21]. Trustworthiness of a Web service is the property of being “worthy of confidence” and therefore related to past consistent reputation in specific context and time.

We define Web service trustworthiness $Cr_t(T_S)$ as a measure of its reputation and is regarded as a predictor of its future behavior [1]. Reputation is what is generally said or believed about a person's or thing's character or standing [14]. It is a collective measure of the opinion of a community of users (humans or agents) regarding their

actual experience with the service [22]. It is computed as an aggregation of users' feedbacks and reflects the reliability and trustworthiness of the service and its provider. Web service reputation is impacted by the following factors [23]:

1. Customer feedback: represents the extent of customer satisfaction from providers' performance based on the interaction with the Web service, and the opinion of the customer on the fulfillment of the service considering the agreement [22] between the user and the service provider.

2. Credibility of a rater: indicates how credible the rater is in providing feedback. Malik and Bouguettaya [24] define a credible rater as one who has performed consistently, accurately, and has proven to be useful (in terms of ratings provided) over a period of time. Ratings from highly credible raters weigh more than ratings from consumers with low credibility's.

3. Customer preference weight: each customer has a specific preference weight for each QoS attribute j denoted by W_i^j in the range [0,1]. Reputation of attribute j at time t denoted by $REP(S_t^j)$; is the weighted average of all feedbacks from all customers N who rated attribute j . For the (j^{th}) attribute, reputation in time (t) can be defined as:

$$REP(S_t^j) = \frac{\sum_{i=1}^N FEEDBACK(S_i^j) * Cr_i^j * W_i^j}{N * W_a^j * Cr_a^j} \quad (2)$$

where $FEEDBACK(S_i^j)$ is received about attribute j from the rater i in the range [0, 1] at time t , Cr_i^j is the rater i credibility in the range [0, 1]. For the (j^{th}) attribute: N^j , W_a^j and Cr_a^j represent number of customers who rated attribute j , average of user preference weights and average raters' credibility respectively.

Web service Global Reputation is the aggregation of all attributes' reputation of the Web service, and defined as:

$$REP(S_t) = \frac{\sum_{j=1}^n REP(S_t^j) * W_a^j}{\sum_{j=1}^n W_a^j} \quad (3)$$

where n is the total number of Web service attributes and (t) is the time stamp. W_a^j is the average of user preference weights for the j^{th} attribute. We model Web service credibility from Trustworthiness component $Cr_t(T_S)$ as:

$$Cr_t(T_S) = \frac{\sum_{j=1}^n REP(S_t^j) * W_a^j}{\sum_{j=1}^n W_a^j} \quad (4)$$

2) Web service Credibility from Expertise component

Expertise, a key dimension of Web service credibility is defined as the degree of a Web service competency to provide accurate results as promised and exhibit high activity [25]. The expertise dimension captures the perceived interoperability and skills of the Web service. QoS monitoring for Web services described in Zeng, Lei et

al. [26] can be used as a reference model. We model Web service credibility drawn from its expertise component as:

$$Cr_t(E_S) = \frac{N_t^S}{N_{Max}^t} * P_s^t \quad (5)$$

where (N_t^S) refers to engagement frequency in a specific period t , and defined as the number of times the Web service was engaged in an execution process. (N_{Max}^t) is the maximum service frequency in that domain; considered as a reference point. (P_s^t) is the performance of service [0, 1]; and computed as the aggregation of all QoS performance. Considering that a quality management system provides temporal information about each attribute performance (P_s^j), i.e., the extent of the service meet the SLA between the user and the provider for that attribute; then we define QoS attribute performance from one transaction P_s^j for the (j^{th}) attribute as follows:

$$P_s^j = \begin{cases} 1 & \text{if } Q_s^{jAdvertised} \leq Q_s^{jPerceived} \text{ (Maximize attribute } j) \\ 1 - \frac{|Q_s^{jAdvertised} - Q_s^{jPerceived}|}{Q_s^{jAdvertised}}, & \text{Maximize } j \text{ Otherwise} \\ 1 & \text{if } Q_s^{jAdvertised} \geq Q_s^{jPerceived} \text{ (Minimize attribute } j) \\ 1 - \frac{|Q_s^{jAdvertised} - Q_s^{jPerceived}|}{Q_s^{jPerceived}}, & \text{Minimize } j \text{ Otherwise} \end{cases} \quad (6)$$

where ($Q_s^{jAdvertised}$, $Q_s^{jPerceived}$) in the range [0, 1] and refer to the advertised and perceived quality values respectively. When the QoS attribute is maximized, means the higher value over the promised (advertised) value is the better, such as security. When the attribute is to be minimized, means the lower value below the promised (advertised) value is the better such as response time and duration. For example, if the advertised response time which needs to be minimized; is (0.8 ms) and the perceived response time is (0.95 ms), then the performance of the response time is (0.8125). While when the perceived response time is (0.75 ms), then the performance of the response time is (1).

Taking the average performance of each attribute from its (N) previous performances as:

$$P_{s,t}^{jAvg} = \frac{\sum_{j=1}^N P_s^j}{N_t^S} \quad (7)$$

Then over-all performance of the service is the weighted mean of all attributes, formally given by:

$$P_{s,t}^t = \frac{\sum_{t,j=1}^n P_{s,t}^{jAvg} * W_a^j}{\sum_{t,j=1}^n W_a^j} \quad (8)$$

where (n) is number of QoS attributes, (W_a^j) is the average preference weight of all users for the j^{th} attribute for all services in that domain over time t .

Using equations (7, 8) in equation (5) this yields expertise credibility at any point of time as:

$$Cr_t(E_S) = \frac{N_t^S}{N_{Max}^t} * \frac{\sum_{t,j=1}^n P_s^{jAvg} * W_a^j}{\sum_{t,j=1}^n W_a^j} \quad (9)$$

3) Computing Web service Credibility

Web service credibility is computed by aggregating the credibility components: trustworthiness component from reputation and expertise credibility component; hence Web service credibility at current time (t) is given by:

$$Cr_t(S) = \beta * Cr_t(T_S) + (1 - \beta) * Cr_t(E_S) \quad (10)$$

where β in the range [0, 1], represents the importance of each credibility component. For example; when ($\beta < 0.5$) the system relies on trustworthiness less than expertise credibility component.

4) Credibility Decay

In Web service selection; recent credibility components: trustworthiness and expertise attract more importance than old ones; considering the decay factor $f_d(t)$ to control this impact; credibility of service (s) can be defined as:

$$Cr(s) = \frac{\sum_{t=t_1}^{t_2} Cr_t * f_d(t)}{\sum_{t=t_1}^{t_2} f_d(t)} \quad (11)$$

where $f_d(t) = e^{-\lambda_1(t_2-t_1)}$, and λ_1 in the range [0, 1], ($t_2 - t_1$) is the time interval difference between the present time and the time in which the credibility data were collected.

V. PERCEIVED RISK AND RISK ATTITUDE IN WEB SERVICE SELECTION

Since no Web service is risk free, there is always some degree of risk or uncertainty associated with Web service selection decisions. In the following section we explore the perceived risk of Web service performance and show how customers have varied risk attitudes towards handling the perceived risk.

A. Perceived Performance Risk in Web service selection

During Web service selection consumers often act on information that is incomplete and far from perfect [3]. As a result, they are often faced with some degree of risk or uncertainty in their selection decisions. Kim, Ferrin et al. [3] formally define perceived risk as a consumer's belief about the potential uncertain negative outcomes from the online transaction. Featherman and Pavlou [27] view perceived risk as "a combination of uncertainty plus seriousness of outcome involved". Perceived risk is commonly viewed as uncertainty regarding possible negative consequences of using a Web service.

In Web service selection, perceived risk has different dimensions such as reliability, availability, response time,

security and privacy; we refer to these dimensions as performance risk. When the service provider does not respect the SLA in any of advertised QoS attributes the Web service performance suffers from such behavior; which in turn increases the severity of the associated risk. For example, when a consumer submits credit card information through a transaction she can feel the threat of the possibility of credit card fraud or even disclosure of consumer information to non-authorized people when the security or privacy performance is low or unknown.

In this paper, we follow [28] and define perceived Performance Risk (PR) in [0,1] as: Consumer assessment of potential performance problems, malfunctioning, transaction processing errors, reliability and/or security problems, that cause the Web service not perform as expected.

B. Risk Attitude and Perceived Risk

Risk attitude represents how willing the customer is to take on the perceived risk which is largely dependent on the character of an individual [21] and their position, e.g., financial position or their role such as followers or leaders. Different factors affect risk attitude such as personality type, gender, age, culture, etc. Furthermore, we believe that risk attitude is context based; for example, a customer can use a Web service without any monetary transaction or even a cheap service with a high attitude to accept the risk, while when using a monetary Web service with payment she would usually have different trade-offs between utility and perceived risk in making her decision.

Consumer risk attitude determines the courses of action to be followed. Consumers who are cautious by nature may avoid risky situations and fail to capture opportunities as a consequence. Since all decisions have an element of uncertainty about them, all decision-makers are risk takers [29]. The degree to which decision-makers enjoy taking risk depends upon individual attitudes.

The risk attitude of the customer plays a vital role in selecting the most attractive choice. However, in Web service selection users may have different risk attitudes; the risk attitude (RA) of a customer is given by a real number in [0, 1]. Customers with risk attitude 0 are the most risk-averse customers, while customers with risk attitude being 1 are the most risk-seeking customers.

C. Perceived Risk from Risk Attitude Perspective

Risk evaluation involves the consumer determining the possibility of the failure of the interaction with the Web service and the subsequent possible consequences for their resources involved in the interaction. In general, it is accepted that the higher the perceived risk the lower the likelihood of the transaction. We believe that credibility, perceived risk and expected utility of the Web service from risk-averse customer perspective are related according to the following the axioms:

1. When WS credibility goes to zero, perceived risk goes to one consequently utility goes to zero.

2. When WS credibility increases, perceived risk decreases and consequently utility increases.
3. When WS credibility goes to one, perceived risk approaches zero, consequently utility goes to the maximum value depending on the customer risk attitude [0, 1].

To model the relation between service credibility, perceived risk and risk attitude, we propose the following formula that satisfies the above axioms:

$$PR(s) = e^{-\mu Cr} \tag{12}$$

where μ is customer risk attitude coefficient in the range [1, 5] and given by $\mu = 4RA + 1$. For a risk-averse customer with risk attitude $RA = 0, \mu = 1$; while for a risk-seeker with $RA = 1, \mu = 5$.

In [30], Sitkin and Weingart (1995) argue that the higher the perceived risk, the greater the perceived chance of experiencing a loss, therefore, the lower the consumer's expected utility from the transaction. Thus we can model the relation between perceived performance risks (PR) from a Web service (S), and associated utility U(s) as:

$$PR(s) + U(s) = 1 \tag{13}$$

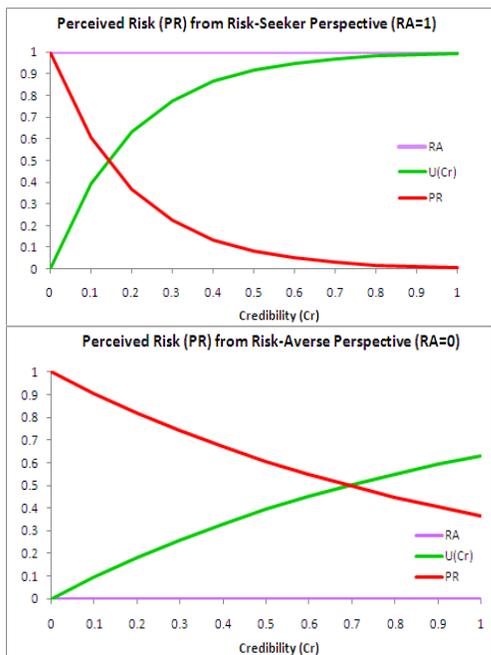


Figure 3. Perceived Risk variation with Different risk attitudes

Fig. 3 shows how the perceived risk (PR) related to the credibility and varies with risk attitude (RA) for the following cases: (1) $RA = 1$ for risk seeker, (2) $RA = 0$ for risk-averse customer. From the above formula we conclude that risk-seeker customers gain more utility than risk-averse customers as shown in Fig. 3, consequently risk-averse customers perceive more risk than risk-seeker customers.

In summary, we believe that the perceived risk is a reflection of user risk attitude i.e., how much risk is the customer ready to take as shown in Fig. 3. For example, if the credibility of the service = 0.7, then from a risk-averse customer perspective with risk attitude = 0 the perceived risk is 0.5, while from a risk-seeking customer perspective with risk attitude = 1 the perceived risk is 0.03.

VI. SERVICE SELECTION WITH RISK ATTITUDE AND PERCEIVED RISK

Customer self confidence assessment is the final determinant in the selection decision process. We argue that customer risk attitude enrich customer confidence-when customer risk attitude increases then customer confidence increases and when customer risk attitude decreases the customer's confidence decreases. The following scenarios describe different customers' behavior in service selection:

1. Risk-seekers customers (Leaders): select the service that maximizes their utility based on Web service credibility and accepting the perceived risk; they usually select the service with the highest credibility score when the perceived risk is within the customer risk attitude. Risk-seeking customers may adopt new services that have never been used before, or they can use a service that they know the perceived risk is high because they have a high risk attitude and choose to accept the perceived risk in order to gain higher utility.
2. Risk-averse customers (Followers): benefit from their social relations and their trust in others, they usually prefer to use a service even if it is expensive it was used by other friends with a proven successful performance. Risk-averse customers usually like to avoid risky situations; they prefer to mitigate the risk by following other trustworthy advice from leaders or other friends than acting themselves.
3. Risk-neutral customers make their decisions based on their risk attitude and the perceived risk from a Web services in hand. They make their decision either to follow other friends to mitigate a high perceived risk or acting as independents if they are confident that they can accept the perceived risk from the transaction.

VII. SIMULATION AND EVALUATION

To demonstrate the feasibility and effectiveness of "Follow the Leader" as a new approach to alleviate the risk in service selection, first we developed a Social Network Analysis Studio (SNAS) using NetLogo platform [31] that analyze user and Web service behaviors in a social network based on our simulation tool "4S: Service Selection Simulation Studio" [32] inspired by Goldbaum (2008). User interface is shown in Fig. 4. We use it to evaluate the validity of our approach. In the following sections we outline the testing environment and outcomes.

A. Simulation Model

Our simulation model is composed of a fixed number of atomic services (9) with the same functional properties and varied in their QoS attributes. Each atomic service maintains a list of QoS attributes and promised values, where QoS is static during any simulation session. Web service credibility is dynamic and computed after each round.

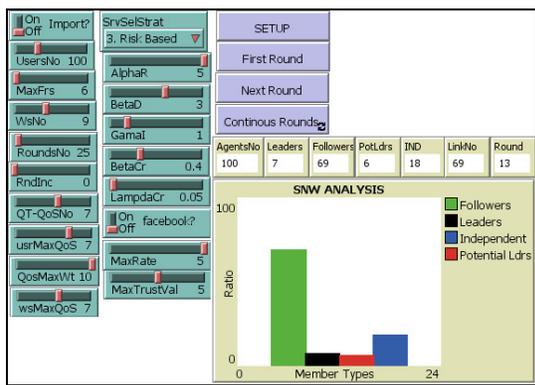


Figure 4. SNA Simulation Tool – User Interface

Each simulation session is composed of a fixed set of rounds (25). Each round represents a time unit e.g., one day. In each round a fixed number of customers (100) enter his/her queries into the system. Each customer has a varied list of preferences and corresponding values and weights. Each customer has a random number of friends (1-6) with corresponding trust values. By the end of each transaction, the system implements credibility computation based on service performance. Each service has an initial credibility at the beginning of each session based on its capabilities.

Each simulation session starts by importing the services and setting customers with their corresponding information. In each round every customer passes its query to the system. The system identifies leaders based on their credibility level and expressive queries. If the customer qualifies as a leader, then the system enables the leader selecting the best service from available services based on the expected utility. If the customer acts as a follower, then the system either: (1) Selects the best friend with highest credibility from the customer’s friends, i.e., the confidence in that friend is higher than the confidence in herself, or (2) Allows the customer to act as independent if the confidence in herself is higher than any of her friends. By the end of each round, each customer provides a feedback to the system about their satisfaction from the service; this feedback is used to derive service reputation which has impact on service credibility.

B. Simulation Results

To test the hypothesis that using the “Follow the Leader” approach is an applicable approach to mitigate the perceived risk in the service selection we perform the following experiments:

1. Impact of Trustworthiness and Expertise on WS Credibility: in this experiment we show how Web service credibility varies with Trustworthiness and Expertise credibility components over time. Fig. 5 shows how credibility components Trustworthiness CR(T) and Expertise CR(E) vary with time, with importance weight ($\beta = 0.4$) for CR(T) and ($1 - \beta = 0.6$) for CR(E) to give CR(Global) for each round.

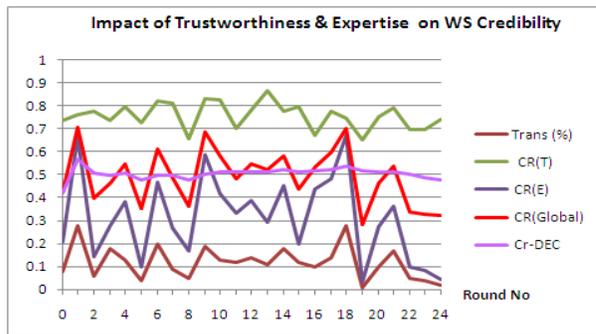


Figure 5. Impact of Trustworthiness and Expertise on WS Credibility on WS (S04)

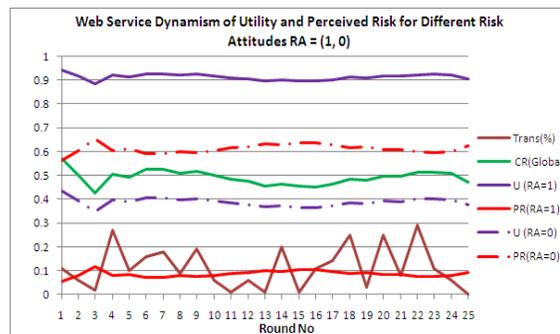


Figure 6. Dynamism of utility and Perceived Risk for Different Risk Attitudes RA = (1, 0)

2. Utility based Credibility vs. Perceived Risk for Different Customers Risk Attitudes RA = (1, 0.5, 0); in this experiment we show how different customers with varied risk attitudes perceive the risk PR from 9 services. In Fig. 6, WS=1 shows the highest credibility (0.73). From a risk-seeker perspective with RA = 1, the perceived risk PR is the lowest (0.026) with the highest utility (0.974); while from a risk-averse perspective with RA = 0, the perceived risk PR is the highest (0.481) with the lowest utility (0.519). This emphasizes the relationship between utility and perceived risk as ($U + PR = 1$), from any customer perspective (i.e., when the utility increases the perceived risk decreases) and vice versa.
3. Dynamism of Utility and Perceived Risk for Different Risk Attitudes RA = (1,0); Fig. 6 shows how credibility and corresponding utility of Web service varies with time from a customer varied perspective (i.e., with risk

attitude as risk-seeker RA=1 and risk averse customer with risk attitude RA=0).

- Malicious Web service behavior – (Facebook’ Privacy Scenario): In this experiment we simulate malicious service behavior after its approved credibility over a specific period of time (first 11 rounds) then acts maliciously by performing inadequately with one of its QoS such as privacy issue [33] for Facebook users. Fig. 7 shows how the service behaves consistently in the first 11 rounds with the highest credibility overall other services, but when one attribute of its QoS suffers, then associated credibility suffers as well. By calculating the impact of this change, we note that Round Credibility (RND-CR) decreased from an average of (0.59) in the first 11 rounds to an average of (0.35) in the rest of simulation rounds, with overall loss in its credibility of (39%). These figures reflect the sensitivity of the model against malicious behavior of Web service.

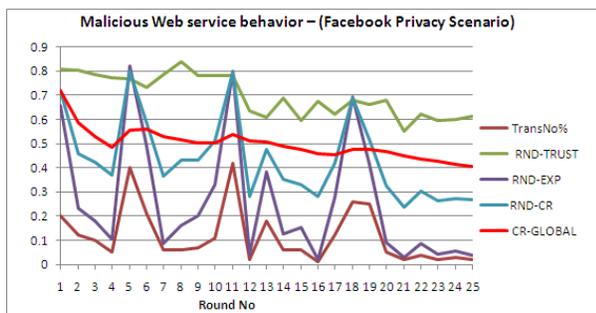


Figure 7. Malicious WS behavior: (Facebook’ Privacy Scenario)

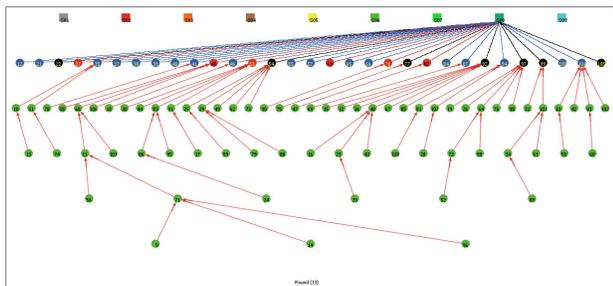


Figure 8. Service Selection based on customer Risk attitude and Service credibility –Follow the Leader Model

- In Web service selection with risk considerations as shown in Fig. 8, Leaders (Black and Red agents) make their selection choice based on their risk attitude and Web service utility. Since leaders risk attitudes are high, they select the service with the highest credibility, whereas for customers with low risk attitude they make their decision based on the confidence that one of their friends selected a high utility service to follow (Green agent). If their confidence in themselves is higher than any of their friends then they take the risk and act as independent (Blue agents). Consequently they select their best service based on service credibility.

C. Results Summary

We summarize our observations from the previous experiments as follows:

- In a Web based social network (WBSN), customer credibility is the determinant of its behavior. Credibility of a customer in a specific domain/context is the predictor of her risk attitude. Usually customers with high credibility act as leaders, while customers with lowest credibility act as followers.
- Web service credibility is the determinant of its behavior; different services in a specific domain have same functionalities and vary in their QoS attributes. Each service has its unique credibility computed based on trustworthiness and expertise. Trustworthiness component is drawn from its reputation while its expertise represents to what extent the service provides promised QoS according to the SLA.
- Proposed Web service credibility model shows its sensitivity to Trustworthiness and expertise. Web service Credibility drops significantly when one or more of its QoS attributes behave maliciously.
- Proposed Web service selection based on risk attitude approach is an efficient approach to alleviate the risk of Web service selection for customers with low risk attitudes i.e., followers. This approach explores the confidence relation between the follower and her friends which is a function of customer credibility.

VIII. CONCLUSIONS AND FUTURE WORKS

In this paper, we proposed a centralized credibility based framework for users in WBSN and for Web services in a specific domain which are similar in their functionality but vary in their QoS. Credibility in both models is drawn from trustworthiness and expertise components of users and Web services. Users’ credibility is an indicator of their risk attitude and self confidence; while service credibility is an indicator of its consumption.

We showed how risk-averse customers make their decisions in Web service selection and follow the best trustworthy friend in their social network; in order to reduce the perceived risk from the available choice based on “Follow the Leader” approach.

We proved the feasibility of our proposed framework in providing accurate Web service selection through simulation. The results of the experiments included in this paper show the applicability and scalability of the proposed credibility assessment based on “Follow the Leader” Model to mitigate the risk in service selection. We have shown how different users with varied risk attitudes make their decisions in the Web service selection process with the perceived performance risk and utility considerations.

Although we handle the risk for followers with low risk attitude in the service selection by following one of their best friends who selected a service that increase the follower utility, considering the confidence relation as the determinant to which is the best friend to follow, notably the

omission of social influences between WBSN members is a limitation which will be explored in a future study.

REFERENCES

- [1] Z. Malik, I. Akbar, and A. Bouguettaya, "Web Services Reputation Assessment Using a Hidden Markov Model," in: Proceedings of the 7th International Conference on Service Oriented Computing, 2009, pp. 576-591.
- [2] Y. Wang and J. Vassileva, "A review on trust and reputation for Web service selection," in Proceeding of the 1st Int. Workshop on Trust and Reputation Management in Massively Distributed Computing Systems, Toronto, Canada, 2007, pp. 25-25.
- [3] D. Kim, D. Ferrin, and H. Rao, "A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents," *Decision Support Systems*, vol. 44, pp. 544-564, 2008.
- [4] T. Vitvar, et al., "Semantically-enabled service oriented architecture: concepts, technology and application," *Service Oriented Computing and Applications*, vol. 1, pp. 129-154, 2007.
- [5] D. Goldbaum, "Follow the Leader: Simulations on a Dynamic Social Network," UTS Finance and Economics Working Paper No 155, <http://www.business.uts.edu.au/finance/research/wpapers/wp155.pdf>, 2008.
- [6] S. Shekarpour and S. D. Katebi, "Modeling and Evaluation of Trust with an Extension In Semantic Web," *Web Semantics: Science, Services and Agents on the World Wide Web*, 2009.
- [7] A. Gürsel and S. Sen, "Producing timely recommendations from social networks through targeted search," in Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems, Budapest, 2009, pp. 805-812.
- [8] T. DuBois, J. Golbeck, J. Kleint, and A. Srinivasan, "Improving Recommendation Accuracy by Clustering Social Networks with Trust," *Recommender Systems & the Social Web*, 2009.
- [9] G. Liu, Y. Wang, and M. Orgun, "Trust Inference in Complex Trust-oriented Social Networks."
- [10] P. Massa and P. Avesani, "Trust-aware recommender systems," in *ACM Recommender Systems Conference(RecSys)*, USA, 2007, pp. 17-24.
- [11] A. Jøsang and S. Presti, "Analysing the relationship between risk and trust," *Trust Management*, pp. 135-145, 2004.
- [12] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, pp. 618-644, 2007.
- [13] Z. Liang and W. Shi, "PET: A PErsonalized Trust model with reputation and risk evaluation for P2P resource sharing," 2005, pp. 201-210.
- [14] A. Jøsang, "Online reputation systems for the health sector," *electronic Journal of Health Informatics*, vol. 3, 2008.
- [15] C. Castelfranchi and R. Falcone, "Social trust: A cognitive approach," *Trust and deception in virtual societies*, pp. 55-90, 2002.
- [16] D. Ramirez-Cano and J. Pitt, "Follow the Leader: Profiling Agents in an Opinion Formation Model of Dynamic Confidence and Individual Mind-Sets," 2006, pp. 660-667.
- [17] F. Andrade, J. Neves, P. Novais, J. Machado, and A. Abelha, "Legal security and credibility in agent based virtual enterprises," *Collaborative Networks and Their Breeding Environments*, pp. 503-512, 2005.
- [18] J. M. Kouzes and B. Z. Posner, *Credibility: How Leaders Gain and Lose It, Why People Demand It*, Revised Edition: San Francisco, CA: Jossey-Bass, 2003.
- [19] J. Al-Sharawneh and M. Williams, "Credibility-based Social Network Recommendation: Follow the Leader," *ACIS 2010 Proceedings*. Paper 24. <http://aisel.aisnet.org/acis2010/24>, 2010.
- [20] J. Al-Sharawneh and M.-A. WILLIAMS, "Credibility-aware Web-based Social Network Recommender: Follow the Leader," *Recommender Systems and the Social Web*, pp. 1-8, 2010.
- [21] D. K. W. Chiu, H. F. Leung, and K. M. Lam, "On the making of service recommendations: An action theory based on utility, reputation, and risk attitude," *Expert Systems with Applications*, vol. 36, pp. 3293-3301, 2009.
- [22] N. Limam and R. Boutaba, "QoS and reputation-aware service selection," in *NOMS: IEEE/IFIP Network Operations and Management Symposium*, 2008, pp. 403-410.
- [23] J. Al-Sharawneh, M. Williams, and D. Goldbaum, "Web Service Reputation Prediction based on Customer Feedback Forecasting Model " in Proceedings of the Fourteenth IEEE International EDOC 2010-Conference, 2010, pp. 33-40.
- [24] Z. Malik and A. Bouguettaya, "Evaluating rater credibility for reputation assessment of web services," *Lecture Notes in Computer Science*, vol. 4831, pp. 38-49, 2007.
- [25] K. Kwon, J. Cho, and Y. Park, "Multidimensional credibility model for neighbor selection in collaborative recommendation," *Expert Systems with Applications*, vol. 36, pp. 7114-7122, 2009.
- [26] L. Zeng, H. Lei, and H. Chang, "Monitoring the QoS for Web services," *Lecture Notes in Computer Science*, vol. 4749, p. 132, 2007.
- [27] M. Featherman and P. Pavlou, "Predicting e-services adoption: a perceived risk facets perspective," *International Journal of Human-Computer Studies*, vol. 59, pp. 451-474, 2003.
- [28] M. S. Featherman and J. D. Wells, "The intangibility of e-services: effects on perceived risk and acceptance," *SIGMIS Database*, vol. 41, pp. 110-131, 2010.
- [29] L. Bauer and U. o. A. F. o. Extension, *Identifying Risk Attitudes: Faculty of Extension, University of Alberta*, 1994.
- [30] C. Yew, V. Tosic, and H. Lutfiyya, "On integrating trust into business-driven management of web services and their compositions," 2008, pp. 102-104.
- [31] NetLogo, "NetLogo Home Page. On line at: <http://ccl.northwestern.edu/netlogo/>," 2009.
- [32] J. Al-Sharawneh and M.-A. Williams, "ABMS: Agent-Based Modeling and Simulation in Web Service Selection," *Proc. Management and Service Science*, 2009. MASS '09. International Conference, pp. 1-5, 2009.
- [33] WIKIPEDIA. Facebook Privacy Issue, Criticism_of_Facebook [Online]. Available: http://en.wikipedia.org/wiki/Criticism_of_Facebook#May_2010, accessed: Aug, 2010

Development Problems in XML Algebraic Parsing Process

Adriana Georgieva
 Fac. Applied Mathematics and Informatics
 Technical University of Sofia, TU-Sofia
 Sofia, Bulgaria
 e-mail: adig@tu-sofia.bg

Bozhidar Georgiev
 Fac. Computer Systems and Control
 Technical University of Sofia, TU-Sofia
 Sofia, Bulgaria
 e-mail: bgeorgiev@tu-sofia.bg

Abstract - In this paper, are presented some problems and solutions concerning the implementation of proposed algebraic method for XML data processing. The proposed theoretical researches and practical realizations lead to faster XML parsing process. Here is suggested a different point of view about the creation of advanced algebraic parser. This point of view is in tight connection with some popular concepts of the functional programming. Therefore, here proposed nontraditional approach for fast XML navigation using algebraic tools contributes to advanced efforts in the making of an easier user-friendly API for XML transformations. This way, the programmer can avoid the difficulties about the complicated language constructions of XSL, XSLT and XPath languages. The choice of programming language C# is a logical consequence, which follows some previous experiments with other high level programming languages. These activities were carried out by the same authors. The discussed specific search mechanism based on the use of algebraic functions is theoretically and practically faster in comparison with many other well-known XML parsers. Finally, the conclusion is that in this area really exists a possibility for creating new software tools, based on the linear algebra theory, which can completely replace the whole XML navigation and search techniques used for the present by XSLT/XPath.

Keywords - hierarchical XML tree structure, functional programming (FP), XML transformations of semi-structured data, algebraic modeling of XML structures, module-finite algebra, XPath scripting language, XML parser.

I. INTRODUCTION

The main purpose of this paper is the research about the possibilities for application of one nontraditional approach for addressing the components in XML tree. This approach is based on the principles of the functional programming (FP). According to the cited incontrovertible sources [1][12], the functional programming is a specific programming technique tightly connected with function definitions. In practice, the difference between a mathematical function and the notion of a "function" used in imperative programming is that imperative functions can have side effects, changing

the value of already calculated computations. That's why, they lack referential transparency i.e., the same language expression can result in different values at different times [12] depending on the state of the executing program. The formal description of BNF (popular as Backus-Naur Form) template concerning functional programming is: *program ::= set of functions*. Functional programming finds use in real practice through several programming languages like Mathematica (symbolic math), F# in Microsoft .NET and XSLT (XML). Spreadsheets can also be interpreted as FP languages. Actually, in the theory are presented many different ways concerning function description - tables, equations, definitions, etc. In view of the fact that the paradigm FP is a mathematical abstraction rather than a programming language, we lay particular stress on types of functions, which are widely used at present. Higher-order functions are functions that can either take other functions as arguments or return them as results. Higher-order functions are closely related to first-class functions [2]. Higher-order functions and first-class functions both allow functions as arguments and results of other functions [8][9]. The distinction between the two functions is subtle: "higher-order" describes a mathematical concept of functions that operate on other functions, while "first-class" is a computer science term describing programming language entities that have no restriction on their use (thus first-class functions can appear anywhere in the program, including as arguments to other functions and as their return values).

Actually, higher-order functions enable partial application or currying, a technique, in which a function is applied to its arguments one at a time, with each application returning a new function that accepts the next argument. In other words, functional programming is a style of programming that emphasizes the evaluation of expressions, rather than execution of commands. The widespread use of XML prompted the development of appropriate searching and browsing methods for XML documents [4]. The presented paper offers a particular point of view focusing on the building of an algebraic formalism

for navigation over XML hierarchy connected with functional programming theory. With the use of XML query languages, users of XML retrieval systems are able to exploit the structural nature of the data and restrict their search to specific structural elements within an XML documents [3].

Definitely, most paradigms for defining a variety of query languages are based on either way of the two logics widely used in the context of trees – first-order logic (FO), and monadic second-order logic (MSO). MSO extends FO by the quantification and navigation over the sets of nodes. In this paper, we shall consider monadic second-order logic (MSO) as first-order logic extended with “monadic second-order variables” ranging over sets of XML elements. In other words, the query languages for extraction of data from XML documents (XSL, XSLT, etc.) are grounded theoretically on MSO logic. XPath language is rather related to FO logic [5].

This article presents a nontraditional point of view that connects the application of FP principles (as illustration of the declarative style of programming) with here proposed algebraic approach. The purpose of this FP model representation is to make the implementation of advanced linear algebra tools [9] possible for XML data manipulations. In Section I the main functionalities of FP and the basic goals of this paper are shown. Section II describes the conceptual model connected with the proposed algebraic approach for faster search in XML hierarchy. The exhibited in this section results follow previous researches of the authors which have been exposed in [2]. Here are discovered the internal links between some theoretical formulae, which show the possible substitution of complicated language constructions (XPath, XSLT) with the discussed above FP techniques. Section III presents some XML parser architectures and program realizations along with an algebraic search and hierarchy access. Finally, in Section IV the general issues, conclusions, the further researches and some open problems are discussed.

II. ALGEBRAIC APPROACH FOR FASTER SEARCH IN XML HIERARCHY

To avoid the bottleneck, that characterizes the languages XSLT/XPath for XML transformations, there is necessity to accelerate the parser process and node access in common XML hierarchy. This section is dedicated to some possibilities for faster search and navigation over XML hierarchical trees by means of linear algebra tools. The presented formulae can be considered like functions, based on module-finite algebra tools [10]. According to cited researches [8] [9] and as result of proposed theoretical model [7], in this paper is given the unique determination of the physical address of the object O_k^r (object r from

level k) in common hierarchical structure i.e., the number p_k^r by the following way:

$$\begin{aligned} p_k^r &= \sum_{i=1}^{k-1} \alpha_i + a_k^I = \alpha_1 + \alpha_2 + \dots + \alpha_{k-1} + a_k^I \\ &= \alpha_1 \cdot \Phi(h_1) + \alpha_1 \cdot \Phi(h_2) + \dots + \alpha_1 \cdot \Phi(h_{k-1}) + a_k^I = \\ &= \alpha_1 \cdot \sum_{i=1}^{k-1} \Phi(h_i) + a_k^I, \end{aligned} \quad (1)$$

where: - $\Phi(h_1) = c_0 = 1$;

$$\Phi(h_2) = c_0 \cdot c_1 = c_1;$$

$$\Phi(h_3) = c_0 \cdot c_1 \cdot c_2 = c_1 \cdot c_2; \dots;$$

$$\Phi(h_k) = c_0 \cdot c_1 \cdot c_2 \cdot \dots \cdot c_{k-1} = c_1 \cdot c_2 \cdot \dots \cdot c_{k-1}$$

are here defined transformed characteristic elements from the tree;

- $c_0, c_1, c_2, \dots, c_{k-1}$ - the number of children (subordinated elements) of any element from level i to level $i+1$; $c_i \in \mathbb{Z}$; ordinary $c_0 = 1$ and therefore:

$$a_k^I = \{ \dots \{ (a_1 - 1) \cdot c_1 + a_2 - 1 \} \cdot c_2 + \dots + (a_{k-1} - 1) \} c_{k-1} + a_k - 1 \quad (2)$$

Here a_k^I is the code value a_k in the hierarchical level k . The calculations in formula (2) are based on the formal description of the sets of code values of XML nodes components. According to suggested formal algebraic description [2] each of the objects in a real XML hierarchical data structure can be accepted as an element of the corresponding hierarchical structure [9]. For XML physical data design, in this article is chosen one-dimensional address array with codes of all XML database elements from the type:

$E(i_1 \div k_1, i_2 \div k_2, \dots, i_n \div k_n)$, which presents the XML data structure in increasing consistency in the order of the corresponding hierarchical levels. On this base the address of element r from level q in common hierarchy can be determinate by:

$$ADR(r, q) = \sum_{m=1}^{q-1} R_m \cdot D_m + r, \quad (3)$$

where: $R_m = (k_{m-1} - i_{m-1} + 1) \cdot R_{m-1}$ and $R_1 = 1$.

In other hand:

$$D_m = (k_m - i_m + 1) / (k_{m-1} - i_{m-1} + 1)$$

Here D_m is the number of the subordinate elements of level $m-1$ to level m . This formula is valid for cases of the “balanced” hierarchical structures, when the number of the subordinate elements to every element of each level to the next one is constant. That’s why, in dependence of the concrete user applications of database structures, the formula (3) is a program, realized so that every element from each hierarchical level has a different number of subordinate elements compared to the next one. If we denote with $a_1^m, a_2^m, \dots, a_{am}^m$ the code value of the elements from level m in common hierarchy, then the expression $k_m - i_m$ is a dimension of level m in hierarchy for each $m = 1, 2, 3 \dots, n$. This algebraic approach allows comparatively simple search of XML hierarchical data by means of the following types of functions – specification functions and nesting functions. As it was shown in [9], the specification functions comprise three basic manipulations for data handling: specification manipulation on only one level, structural specification manipulation that returns all lower levels and quantity specification level – returns all possible levels in horizontal and vertical order. As can be seen in these specification functions, the existent relationships between the elements of the different hierarchical levels are mainly from two types: relations between elements within the structure – *inside-structure relations* and relations between elements of the different hierarchical structures – *inter-structure relations*. Most of these relationships are either from type “one-to-one” or from type “one-to-many”.

Let us consider two basic hierarchical structures \mathbf{T} and \mathbf{P} in one XML document and corresponding relations between elements of them.

Definition 2.1. The relations between elements from the same hierarchical structure representing relationships between them, as in the one hierarchy - structure \mathbf{T} (or in other hierarchy- structure \mathbf{P}) we call relations of strict order and strict inclusion [7][10].

More complicated are the relations from the type “one-to-many” that present the relationships between elements of some hierarchical levels in the \mathbf{P} -structure – for example the relations between elements of every couple of levels K and L . For these relations is defined the operation “projection” (\mathbf{pr}) for each element from K to L .

Definition 2.2. Each element $k_i \in K$ correlates with non-empty set of elements $\{l_j\} \in L$, which we call “intersection” by k_i and will denote with $\mathbf{r}(k_i, l_j)$.

The intersections by k_i i.e., given element is a set of such subjects $\{l_j\}$, that $(k_i, l_j) \in \mathbf{r}$, where $i=1, \dots, n$; $j=1, \dots, m$ (n is a number of elements of K , m is a number of elements of L and it is not obligatory $i \neq j$). For example, when:

$$K = \{k_1, k_2, k_3, k_4\}, L = \{l_1, l_2, l_3\} \quad \text{and}$$

$$\mathbf{r}\{(k_1, l_1), (k_1, l_3), (k_2, l_1), (k_2, l_2), (k_3, l_2), (k_3, l_3), (k_4, l_4)\}$$

$$\text{then } \mathbf{r}[k_1] = \{l_1, l_3\}; \quad \mathbf{r}[k_2] = \{l_1, l_2\};$$

$$\mathbf{r}[k_3] = \{l_2, l_3\}; \quad \mathbf{r}[k_4] = \{l_3\},$$

moreover always exist at least one $\mathbf{r}[k_i] \neq \emptyset$,

because projection $\mathbf{pr}(r(k_i, _)) \neq \emptyset$ as, for

example, $\mathbf{pr}(r(k_i, _)) = \{l_1, l_3\} \neq \emptyset \dots$, etc.

These intersections by $k_1, \dots, k_n \in K$ present the specific peculiarities of relation \mathbf{r} between elements of levels K and L in hierarchical structure \mathbf{P} . Similarly, here can be described the relations between other couple of hierarchical levels from this type in any XML document.

III. HOW THE PROGRAM SYSTEM (ALGEBRAIC PARSER) IS BUILT?

For the purposes of presented research is assumed that the physical records in the XML hierarchical file are with fixed length. On Fig. 3.1 is depicted the functionality of proposed in this article XML parser. Usually, XML documents are stored in physical memory of computer by means of a variety of index-sequential methods. Each element from a given level in the common hierarchy includes different number of siblings and child elements. It means that any object O_n^r is represented with the code value (integer) p_n^r , which is defined from the disposition of the element in XML hierarchy. Here n is the number of hierarchical level in common structure and r is the place number in the fixed level from left to right. This algebraic processor supports a code table with the names of elements of current XML document and the corresponding integers p_n^r , which uniquely define the place of the object (node) O_n^r from level n in the real hierarchical structure i.e., its address in the physical XML database.

These algebraic presentations of the binary relationships in hierarchical structures remove the necessity from table’s work, relation schemes, etc. So it operates only with rows of numbers, which leads to

use of ordinary algebraic tools for data transformation from XML structures to their presentation on physical level.

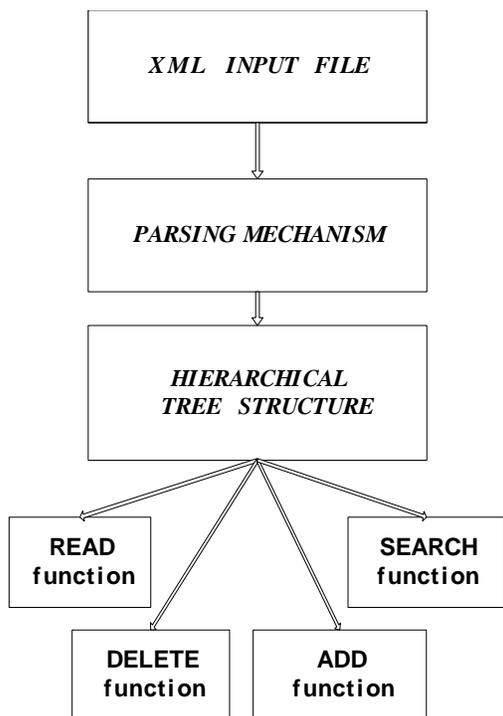


Figure 3.1. Common scheme of functional XML parser

The file information after the parsing processing is saved in the same way as it is kept in the XML input hierarchy. Search function gives the opportunity to the user for tag searching procedure. Final results show the content of the corresponding tag and its position in XML hierarchy. The algebraic approach makes obvious the possibilities for reaching linear time functions in XML tree hierarchy handling. The user can insert an additional information in XML file according to previously defined input format. Some operations for common use could be done as follows:

- DELETE operation of element a_k^f - in this case $a_k^f=0$ means removal of this element from the table along with its descendants down in the hierarchy until to the last level n .
- GET operation uses the formula (1) for immediate address search of an assigned in advance element.
- INSERT operation puts in the table the name of element and its coordinates; here is necessary to increase the index of other elements on the right side of the element, for example a_k^{m+1} , etc.

For more flexibility and best user convenience there is foreseen the possibility that provides WEB access to the parser. The presented diagram on Fig.

3.3. describes in details classes, used in project realization.

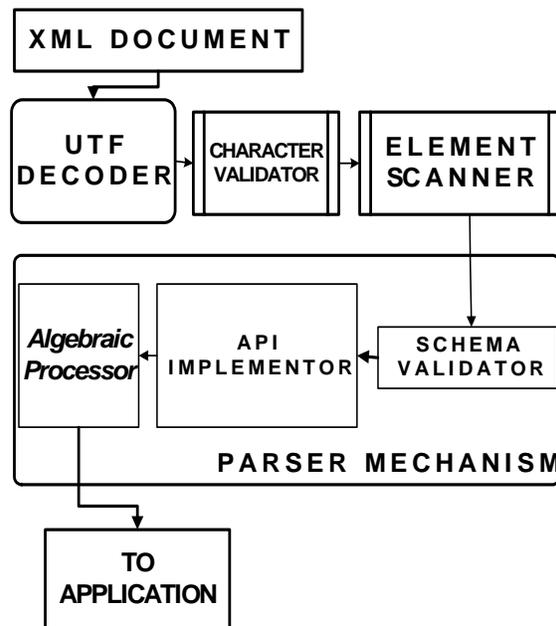


Figure 3.2. XML Parser Architecture added to the algebraic searching and hierarchical tree access

On Fig. 3.2 is defined a set of normative functions for use with the proposed in this paper processor.

For the program realization of XML parser here is used MS Visual Studio environment. Visual Studio [11] is chosen for this purpose in view of the fact that this programming package provides wide spectrum of software tools, used to develop both console applications and WEB applications (and services) as well. Visual Studio supports different versions of programming languages C/C++, VB.NET, C#, XML/XSLT, HTML/XHTML, JavaScript and CSS. Other popular languages can be easily supplemented to this MS program environment after simple installation. The following elements of MS Visual Studio are used in the building process of XML parser:

- Code editor which facilitates text colors and recognizes variables, functions, methods and other components through its core module IntelliSense.
- Debugger module is implemented for tracing programing code about errors detection and correction in input file.
- Designer modules: Windows Forms Designer for creating Windows Forms graphical interface and Web designer/development for WEB sites applications.

The first practical realization of so proposed algebraic parser uses programming tool Eclipse SDK for Java [9]. This article is dedicated to the acceleration process of this parsing mechanism using language C#.

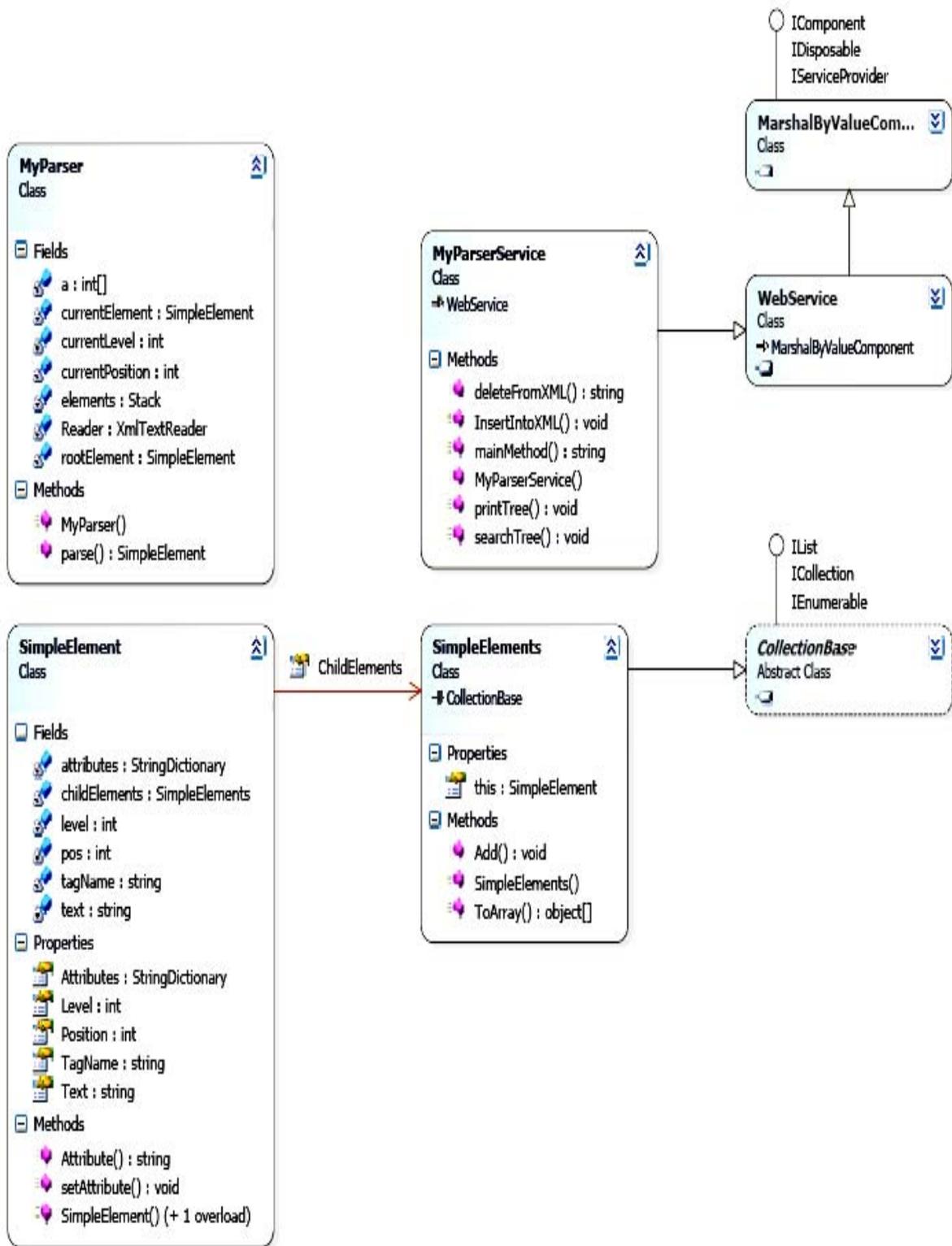


Figure 3.3. Functional class diagram of the parser

The results of both practical researches (JAVA and C# realizations) demonstrate faster accomplishment of discussed XML operations in comparison to some traditional approaches, especially to these which are based on the search in XML hierarchy (XSLT/XPath).

IV. CONCLUSION AND FUTURE WORK

This article is an author's attempt to create a mechanism for accelerating XML document processing, connected with the main principles of FP. The presented XML parser is built and practically works by using advanced algebraic formulae. The authors reveal several basic algebraic operations, which are included in proposed in this article parser and make logical connections with some concepts of the functional programming. On the basis of this model, that presents the addresses of elements in XML hierarchical document as integer values, it is possible to work with common algebraic mechanism. This mechanism is used in modeling of relationships between the different hierarchical components, data operations and standard specification functions [6]. It allows a possibility to work with ordinary linear operators in harmony with FP theory instead of the more difficult operations, which are specific for the widely used models. The algebraic approach points out some opportunities for search and navigation in different WS*-specifications especially as WSDL, BPEL, UDDI, etc. Here are suggested algebraic mechanisms for advanced algebraic processor creation (with all necessary programming modules). This nontraditional (functional) approach about the faster navigation with the presented algebraic tools promotes to build new interface for XML search techniques and transformations. The implementation of so proposed method in the area of SOA could accelerate the various types of XML message-based communications concerned WSDL, UDDI, BPEL, SOAP, etc. This algebraic mechanism can be dynamically located, invoked and combined with other navigation techniques. It gives users the opportunity to process their data messages easier and faster, that is of critical importance to make service-oriented paradigm operational in the real practical environment. Finally, the proposed approach is different in comparison with many other well known query and transformational languages in respect of their definition, expressiveness and search techniques. Several research questions remain as it is mentioned below:

- The results of this paper justify a natural fixed point in the development of some future possibilities for matrix presentation of the relationships between the elements of the different hierarchical levels in XML hierarchy (inside structure and inter-structure relations). Actually, using matrices will be very convenient for conceptual representation and programming realization of hierarchical relationships between every two levels in the whole data structure.

- The presented paper offers an algebraic point of view for building of algebraic processor (with appropriate software and programming techniques). Therefore, this particular point of view about an algebraic processor based on the tools of linear algebra and motivated from FP theory, conducts to software results, which eliminate the complicated language constructions of XSL, XSLT and XPath.

REFERENCES

- [1]. J. Darlington, P. Henderson, and D. A. Turner, Functional programming and its applications, Cambridge university press, ISBN 0 521 24503 6, 1982
- [2]. A.Georgieva and B. Georgiev, "A Navigation over XML Documents through Linear Algebra Tools", The Fourth International Conference on Internet and Web Applications and Services - ICIW 09, Venice/Mestre, Italy, 24-29 May 2009, Published by IEEE Computer Society, ISBN: 978-0-7695-3613-2/09.
- [3]. J. Keogh and K. Davidson , XML DeMYSTiFied, McGraw-Hill, Emeryville, California, USA, 2005.
- [4].World Wide Web Consortium, <http://www.w3.org/TR/2004/> Extensible Markup Language (XML) 1.0. W3C Recommendation, third edition, February 2004.
- [5]. L. Libkin, "Logics for unranked trees: an overview", Department of Computer Science, University of Toronto, 2006.
- [6]. B. Georgiev, "An Approach for Some Implementations of W3C-Standard XML", Second International Scientific Conference, Kavalla, Greece, 2005.
- [7]. A. Georgieva,"Algebraic Modelling of Hierarchical Data Structures on Conceptual Level", Proceedings of Second International Scientific Conference "Computer Science'2005, Chalkidiki, Greece, Part II, pp.113-118.
- [8].A. Georgieva and B. Georgiev," Conceptual Method for Extension of Data Processing Possibilities in XML Hierarchy", Forth International Scientific Conference, Kavalla, Greece, 2008.
- [9]. B.Georgiev and A.Georgieva, "Realization of Algebraic Processor for XML Documents Processing", AIP Conference Proceedings of 36-th International Conference AMEE-2010, vol.1293, pp. 279-286.
- [10]. L. Garding and T. Tambour, "Algebra for Computer Science", Spring-Verlag, N.Y., 1988.
- [11]. J.Sharp, Microsoft Visual C# Step by Step, Microsoft Press, Washington, 2008
- [12]. P. Hudak, "Conception, Evolution, and Application of Functional Programming Languages", ACM Computing Surveys 21/3, 1989, pp. 359-411.

A Privacy Policy Framework for Service Aggregation with P3P

Liju Dong^{*†}, Yi Mu^{*}, Willy Susilo^{*}, Peishun Wang^{*}, Jun Yan[‡]

^{*}Centre for Computer and Information Security Research,

School of Computer Science and Software Engineering, University of Wollongong, Wollongong, NSW 2522, Australia

[†] School of Information Science and Engineering, Shenyang University, Shenyang 110044, P. R. China

[‡] School of Information System and Technology, University of Wollongong, Wollongong, NSW 2522, Australia

Email: {liju,ymu,wsusilo,peishun,jyan}@uow.edu.au

Abstract—Service aggregation has exhibited useful features for efficient and reliable services, especially for the Internet. Recent advances of service aggregation pose a new challenge to privacy policy management due to the nature of policy aggregation and policy inconsistency. Previous studies in privacy policies do not capture privacy issues in service aggregation. In this paper, we present a formal result to demonstrate privacy policy aggregation. In particular, we show how to implement privacy policy aggregation with Platform for Privacy Preferences (P3P).

Keywords: Privacy Policy, Service Aggregation, P3P.

I. INTRODUCTION

The advances of the Web technologies and service oriented architecture enable much better services to Web users in terms of the volume of services and the efficiency of services. It is a trend to combine various services from different providers in order to offer better and efficient services to customers. As an emerging technology, Service Aggregation has been regarded as a promising candidate for integrate services from multiple service providers [1], [2], [3]. Its benefit originates from the added value generated by the possible interactions and by the large scale rather than by the capabilities of its individual service provider separately. This technology has created tremendous opportunities to businesses. On the other hand, it also raises new security issues, as services are provided by service providers from distributed network domains [4]. These issues have not been addressed in the literature.

Privacy is always an important issue in web services [5], [3], [6]. Many organizations now publish their privacy policies on their online service web sites. In a single domain environment, a well-defined privacy policy can be formally presented with the well-known privacy policy languages such as P3P [7], [8] and XACML [9]. However, the situation is entirely different, while the service is provided through an aggregate server, where multiple service providers behind the aggregate server normally adopt different privacy policies and the aggregate service requires an aggregate privacy policy. The major difficulty to policy aggregation is due to inconsistency and conflict of the corresponding policies, where each server provides a part of the service. There exist several other useful tools for privacy policy management, such as APPEL [10], EPAL [11], [12], [13] and ASL [5]. These tools provide formal approaches for describing privacy policies, but they do not capture privacy policies in service aggregation.

There exist several privacy policy models for multiple privacy policies in the literature. As one of the most notable works, Backs *et al.* [11] proposed a formal model for composing

enterprise privacy policies. The aim of the model is to provide the compliance with different privacy policies when several parts of an organization or different enterprises cooperate. This work is based a superset of the syntax and semantics of IBM's Enterprise Privacy Authorization Language (EPAL). They provided an elegant solution to handle conjunction and disjunction of privacy policies, which are not well defined in EPAL. We notice that this policy model does not accommodate our model where the conflicts in privacy policy aggregation possess a more complex nature, which cannot be captured with logical AND and OR defined in their model.

Backes *et al.* [11] proposed a formal model for comparison of enterprise privacy policies in P3P (E-P3P), based on EPAL. Although well-established in the theory, the problem addressed in their work is mainly about how to efficiently check whether one policy refines another. This privacy policy model does not capture all in privacy policy management, especially in conflict resolution in service aggregation. With other novel functionalities, several other privacy policy comparison models were introduced [14], [15], [5], [16], [17]. However, these methods do not address the privacy management for service aggregation either. In particular, they do not consider large privacy policy sets from multiple parties.

In this paper, we formally define the privacy policy aggregation and provide an instantiation with P3P to demonstrate how to implement aggregate privacy policies. In particular, we present the definitions including privacy policy aggregation, privacy policy aggregation with P3P, policy comparison for P3P, and concrete P3P examples. We also provide a solution to privacy policy conflict and constraint.

The remaining sections of this paper are organized as follows. In section II, we present an overview of service aggregation and provide a description about the service aggregation model we consider. In Section III, we preset the syntax and semantics of our privacy policy model. In Section IV, we provide the proposed implement of our framework to P3P. In Section V, we conclude this paper.

II. OVERVIEW OF SERVICE AGGREGATION

Service aggregation is associated with methods and tools that create composite services and their lifecycle management including alignment of privacy policies, security, transaction management, quality of service and other elements of service provision. In the Internet, the service providers are geographically distributed. A service aggregator manages the services

requested by a client. The aggregated service could be one or multiple services in terms of the client's requirement. A general view of a service aggregation system is given in Figure 1.

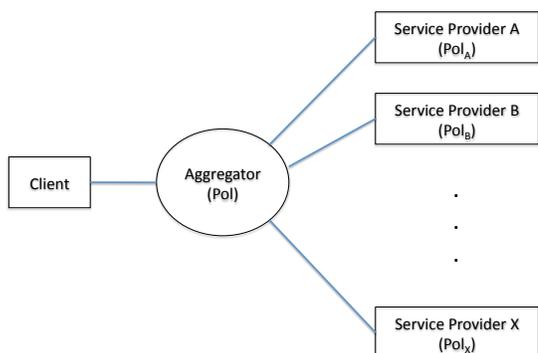


Figure 1. Selected services are aggregated by the Aggregator. Accordingly the related privacy policies $\{Pol_i\}$ are aggregated into Pol .

In general, the service aggregator is the only entity that is visible to the client. To obtain a suitable service, a client needs to register with the aggregator. For getting a service, a client only needs to contact the aggregator, who in turn contacts the related service providers located at the backends. The merit of such type of services is that clients do not need to hunt required services, while they only need to contact the aggregator; therefore, it simplifies the process.

Because of the distributed nature, security and privacy are important issues in service aggregation. Nowadays, Web services are required to publish their privacy policies. In service aggregation, the final service might consist of multiple services whose privacy policies are different. As depicted in Figure 1, only privacy policy that is visible to the client is an aggregated privacy policy from a multiple policies of involved service providers. The aggregated policy must reflect all the policy rulesets from the service providers. It is desirable that the aggregated policy is dynamically formed, as it depends on the related services selected by the service aggregator. To ensure efficiency and accuracy in forming an aggregated policy, in this paper, we develop a privacy aggregation framework along with associated policy aggregate algorithms.

III. SYNTAX AND SEMANTICS

In this section, we formally define privacy policy aggregation, with the consideration of the P3P instantiation. The objective of this section is to provide the reader with clear definitions that illustrate the P3P instantiations presented in the next section. Those definitions are presented with a simple formal language so that it can be easily understood by practitioners.

A. Syntax of Privacy Policies

A privacy policy is a tuple of a vocabulary, a set of authorization rules, and a default ruling. The vocabulary defines subjects, objects, operations, and purposes. Subjects are a set of users.

Objects are data. Our model also includes Purpose, which is an important element for privacy policies.

Definition 1: (Vocabulary) A vocabulary is a tuple $\mathcal{V} = (S, O, OP, P)$ where $S, O, OP,$ and P are elements called subject, object, operation, and purpose.

To define the privacy policies, we require that all components also get a subscript i and an authorization ruling $\mathcal{A} = \{+, -, *\}$, where $\{+, -, *\}$ denote “allow”, “do not allow”, and “do not care” or from user perspective, “require privacy information”, “do not require privacy information”, and “do not specify”.

Definition 2: (Ruleset) A ruleset for a vocabular \mathcal{V} is a subset of $\mathcal{I} \times S \times O \times OP \times P \times \mathcal{A}$, where \mathcal{I} denotes an index set. An instance of a ruleset is denoted by a complete set $rs = (i, s, o, op, p, a)$ or a subset of (i, s, o, op, p, a) .

Definition 3: (Privacy Policy) A privacy policy Pol is a triple $(\mathcal{V}, \mathcal{R}, \mathcal{A})$ of a vocabulary \mathcal{V} , a ruleset \mathcal{R} , and a related ruling \mathcal{A} .

We assume that the components of a privacy policy Pol are always called as in Definition 3. For simplicity, we denote by $Pol = (r_1, r_2, \dots, r_n)$ a privacy policy that consists of a tuple of n rules, where r_i is an instance of $(\mathcal{V}, \mathcal{R}, \mathcal{A})$.

B. Semantics of Privacy Policy Aggregation

Definition 4: (Policy Aggregation) Let $Pol_k = (r_{k,1}, r_{k,2}, \dots, r_{k,n_k})$ for $k = 1, 2, \dots, m$ be m privacy policy sets, where $r_{k,i}$ denotes the corresponding rule. The aggregate privacy policy is defined as

$$Pol(R_I) \leftarrow \bigoplus_{k=1}^m Pol_k(r_{k,i}) \quad (1)$$

where by $Pol(r)$ we denote that rule r belongs to policy Pol , by $r_{k,i}$ we denote a rule that is indexed by i and belongs to policy k , by R_I we denote the resulting rules and by \bigoplus we denote a generic aggregate operator representing Union, Intersect, Minus, or Conflict, depending upon the properties of the policy rules. ‘ \leftarrow ’ denotes an assignment, where the policies on the right are aggregated to yield the policy on the left.

As an example of two policy sets, we have

$$Pol(R) \leftarrow Pol_1(a) \oplus Pol_2(b)$$

where a, b are two rules and R is the aggregate rule set, which could be $a, b, (a, b)$, or an empty set. The “rule” is a generic term that can be an element, a statement, or a constraint in P3P.

Definition 5: (related) Two rules a and b are said related (or $a \sim b$), if they are associated with the same privacy property. If two rules are non-related (or $a \not\sim b$), then the aggregate rule is an union of them; that is,

$$Pol(a, b) \leftarrow Pol_1(a) \oplus Pol_2(b).$$

As an example, if both a and b are associated with gender, they are regarded as “related”. Otherwise, if a is about gender and b is about driving license, then they are “non-related”.

Definition 6: (privacy-expose) A rule is said privacy-expose, if the rule queries a piece of private information; otherwise, the rule is non-privacy-expose. A privacy-expose rule overrides an non-privacy-expose rule if they are in conflict and related.

For example, a_i is a policy item requesting a piece of personal information (privacy-expose) and b_j does not require it. In the view of the customer, he only cares about his privacy before he commits to the aggregate service.

The rules in a privacy policy are subject to the following operations. ‘=’ is an operator indicating that the left and the right are equivalent. ‘>’ is an operator indicating the left overrides the right. ‘<>’ is an operator indicating that the left is in conflict with the right. They are captured by the following three definitions.

Definition 7: (equivalent) If two policy rules a and b represent the same privacy policy, then they are equivalent; that is $a = b$.

Definition 8: (overrides) If a policy rule a dominates another rule b , a overrides b (or $a > b$).

It is dependent on the condition that a rule is allowed to override the other. For example, a privacy-expose rule can override a non-privacy-expose rule. To deal with the conflict of policies, we introduce the conflict resolution definitions.

Definition 9: (in conflict) Given two rules a and b , if neither rule overrides the other, then they are in conflict or $a <> b$.

In the following, we demonstrate how to implement aggregation by using some typical examples.

Example 1. If $a_i \in Pol_1$ and $b_j \in Pol_2$ are not equivalent ($a_i \neq b_j$) and are unrelated ($a_i \not\sim b_j$), then we have

$$Pol(a_i, b_j) \leftarrow Pol_1(a_i) \oplus Pol_2(b_j).$$

In this instance, $Pol \leftarrow Pol_1 \cup Pol_2$. Here, we have omitted other rules in the policy sets for simplicity.

Example 2. If $a_i \in Pol_1$ and $b_j \in Pol_2$ are equivalent, then $a_i = b_j$. We have

$$\begin{aligned} Pol(a_i) &\leftarrow Pol_1(a_i) \oplus Pol_2(b_j) \\ \text{or } Pol(b_j) &\leftarrow Pol_1(a_i) \oplus Pol_2(b_j). \end{aligned}$$

In this instance, $Pol \leftarrow Pol_1 \setminus Pol_2$ or $Pol \leftarrow Pol_2 \setminus Pol_1$, where “\” denotes “exclude”.

Example 3. If $a_i \in Pol_1$ is a privacy-expose rule and $b_j \in Pol_2$ is a non-privacy-expose rule, with $a_i \sim b_j$, then $a_i > b_j$. We have

$$Pol(a_i) \leftarrow Pol_1(a_i) \oplus Pol_2(b_j).$$

In this instance, $Pol \leftarrow Pol_1 \setminus Pol_2$

Example 4. $a_i \in Pol_1$ and $b_j \in Pol_2$, with $a_i \sim b_j$. $a_i <> b_j$, if there is no any conflict resolution. We have

$$Pol() \leftarrow Pol_1(a_i) \oplus Pol_2(b_j).$$

In this instance, $Pol \leftarrow Pol_1 \cap Pol_2$.

Definition 10: (Semantics of Policy Aggregation). Given privacy policy Pol_k , the evaluation result of Pol_k is defined by the following algorithm:

1. Select as input two policy rules a_i and b_j from two policy sets: $Pol_1(a_1, \dots, a_{n_1})$ and $Pol_2(b_1, \dots, b_{n_2})$.
2. Compare the selected rules in terms of the definitions 5-9 and output the aggregate policy.

3. Repeat the process till the all rules in the target policy sets are checked.

The aggregation algorithm is referred to as Algorithm 1 described in Figure 2 in detail.

```

input  $n_1, n_2$ 
while  $i < n_1$  and  $j < n_2$ , do
  input:  $Pol_1(a_i), Pol_2(b_j)$ 
  if  $a_i = b_j$  then
    return  $Pol(a_i)$  or  $Pol(b_j)$ ;
  end if
  if  $a_i \not\sim b_j$  then
    return  $Pol(a_i, b_j)$ ;
  else
    if  $a_i <> b_j$  then
      return  $Pol()$ ;
    else
      if  $a_i > b_j$  then
        return  $Pol(a_i)$ ;
      else
        return  $Pol(b_j)$ ;
      end if
    end if
  end if
end
    
```

Figure 2. Algorithm 1: an algorithm of privacy policy aggregation.

IV. P3P POLICY IMPLEMENTATION IN SA

A. P3P Deployment

A common way to express privacy principles are privacy policies expressed in formal implementable languages, such as P3P (the Platform for Privacy Preferences) [18], XACML [19] and some other languages. P3P is the most popular policy language since 2006, which is an industry-supported self-regulation approach to privacy protection [20]. It is a W3C recommendation as a protocol to communicate how a service intends to collect, use, and share personal information about its visitors [21], [22]. The current development status of P3P is the Working Group Note of the P3P 1.1 Specification, published in November 2006. P3P is an industry standard for privacy protection, designed to give users more control over their personal information when visiting services. We choose P3P as an example for the privacy policy implementation in SA.

A policy set $Pol(a_1, \dots, a_n)$ is represented by a P3P statement. A rule in a P3P policy set can be represented by an element or a constraint in P3P. According to this definition, a P3P policy may consist of several policy sets:

$$\{Pol_1(a_1, \dots, a_{n_1}), Pol_2(b_1, \dots, b_{n_2}), \dots, Pol_m(x_1, \dots, x_{n_m})\}.$$

To clarify our definition, we consider the following example. There are several elements in one statement, such as Purpose, Retention, Recipient, Data and other options. For example, Pol_i is a statement as below:

```

<STATEMENT>
  <PURPOSE><current/><develop/></PURPOSE>
    
```

```
<RECIPIENT><ours/><delivery/></RECIPIENT>
<RETENTION><indefinitely/></RETENTION>
<DATA-GROUP>
  <DATA ref="#thirdparty.name"/>
  <DATA ref="#thirdparty.home-info"/>
  <DATA ref="#thirdparty.business-info"/>
</DATA-GROUP>
</STATEMENT>
```

The statement is referred to as a policy set:

$$Pol_i(\text{PURPOSE}, \text{RECIPIENT}, \text{RETENTION}, \text{DATA-GROUP}).$$

Obviously, Pol_i does not reflect the entire policy, as there are multiple layers in P3P. In the following section, we present a solution by considering the entire P3P setting.

B. P3P Implementation

To implement P3P policies, we classify a P3P statement into levels in terms of depth.

Definition 11: (Element Set (ES)) An Element Set consists of P3P elements, which can be normal elements and optional elements. Optional elements have an optional value 0, 1, 2, or 3, which denote none, always (as default), opt-in, and opt-out, respectively. These elements are arranged with levels (ESL): 0, 1, ..., n , in terms of depth, where 0 is the root and n is the last ESL or leaves.

Taking the above P3P statement as an example, we have

- ESL = 0: <STATEMENT>.
- ESL = 1: <PURPOSE>.
- ESL = 2: <current/><develop/>.
- ESL = 1: <RECIPIENT>.
- ESL = 2: <ours/><delivery/>.
- ESL = 1: <RETENTION>.
- ESL = 2: <indefinitely/>.
- ESL = 1: <DATA-GROUP>.
- ESL = 2: <DATA ref="#thirdparty.name"/>,
 <DATA ref="#thirdparty.home-info"/>,
 <DATA ref="#thirdparty.business-info"/>.

In the last P3P example, the statement has a depth of 2. The root level (ESL = 0) is <STATEMENT>. The first level (ESL = 1) contains a set of default tags such as <PURPOSE>, <RETENTION>, <DATA-GROUP>, etc. The second level (ESL = 2) contains a number of elements depending on their parent. As shown in Table I, <PURPOSE> includes a set of children, which could be either optional or non-optional.

The aggregate Algorithm 2 along with Algorithm 3 and Algorithm 4, as shown in Figure 3, can be utilized to achieve an aggregate privacy policy. It illustrates how two sets of privacy policies can be aggregated. The algorithm can be extended to more policy sets, when the input is altered.

C. P3P Example

To illustrate our privacy policy aggregation algorithms, we provide a concrete P3P example. Assume that $Pol_1(\text{statement})$ and $Pol_2(\text{statement})$ are privacy policies of two online bookshops, respectively.

Table I

THE THIRD COLUMN LISTS THE P3P ELEMENTS OF <PURPOSE>, WHICH CAN BE EITHER OPTIONAL OR NON-OPTIONAL WITH OPTIONAL VALUES GIVEN IN THE FIRST COLUMN. AS AN EXAMPLE, <CONTACT/> IN THE SECOND COLUMN IS USED TO COMPARE WITH THE ELEMENTS LISTED IN THE THIRD COLUMN. AS A RESULT, THEY COULD BE EITHER RELATED OR NON-RELATED, AS DEFINED IN SECTION 2.

	ES (ESL = 2)	ES (ESL = 2)	Related
Optional Value	<contact/>	<current/>	no
(Ovalue):	<contact/>	<admin/>	no
	<contact/>	<develop/>	no
	<contact/>	<tailoring/>	no
0 none	<contact/>	<pseudo-analysis/>	no
1 always	<contact/>	<pseudo-decision/>	no
2 opt-in	<contact/>	<individual-analysis/>	no
3 opt-out	<contact/>	<individual-decision/>	no
	<contact/>	<contact/>	yes
	<contact/>	<historical/>	no
	<contact/>	<telemarketing/>	no
	<contact/>	<other-purpose/>	no

$Pol_1(\text{statement})$ says that the name, postal address, and miscellaneous online data are used for completing the current data transaction. P3P policies collect personal information only for the current service. Considering the attributes of the purpose opt-in and opt-out values, we can simplify them as <contact required=opt-out/>. $Pol_1(\text{statement})$ also uses users' data history and offers personalized book recommendations by categories <preference/>.

$Pol_1(\text{statement}) = Pol_1(\text{purpose}, \text{recipient}, \text{retention}, \text{data-group})$ represents the following P3P privacy policy.

```
<STATEMENT>
  <PURPOSE>
    <current/><admin/>
    <contact required="opt-out">
  </PURPOSE>
  <RECIPIENT><ours/></RECIPIENT>
  <RETENTION><stated-purpose/></RETENTION>
  <DATA-GROUP>
    <DATA ref="#user.name"/>
    <DATA ref="#user.home-info.postal"/>
    <DATA ref="#dynamic.miscdata">
      <CATEGORIES><online/>
    </CATEGORIES>
  </DATA>
  <DATA ref="#dynamic.miscdata">
    <CATEGORIES><preference/>
  </CATEGORIES>
  </DATA>
</DATA-GROUP>
</STATEMENT>
```

$Pol_2(\text{statement}) = Pol_2(\text{purpose}, \text{recipient}, \text{retention}, \text{data-group})$, as given below, says that $Pol_2(\text{statement})$ requires to use the miscellaneous purchase data to create personal recommendations, where the user name and miscellaneous purchase data will be used for the current purchase transaction.

```
<STATEMENT>
  <PURPOSE>
    <current/><admin/>
    <contact required="opt-in"/ >
  </PURPOSE>
  <RECIPIENT><ours/></RECIPIENT>
```

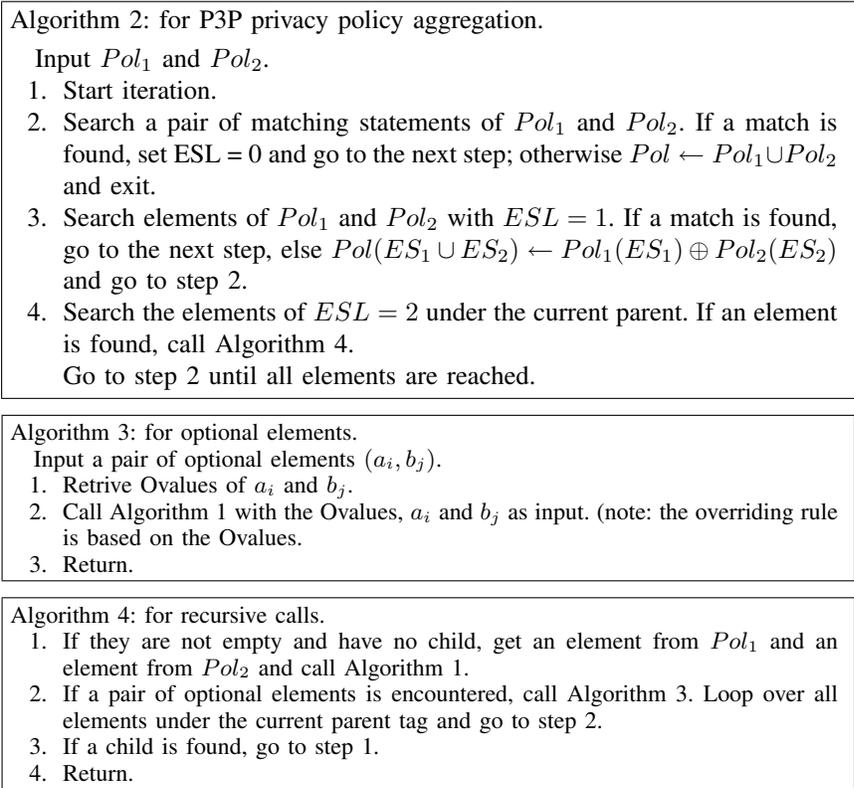


Figure 3. P3P aggregation algorithms.

```

<RETENTION><indefinitely/></RETENTION>
<DATA-GROUP>
  <DATA ref="#user.name"/>
  <DATA ref="#dynamic.miscdata"
    optional="yes">
    <CATEGORIES><content/>
  </CATEGORIES>
</DATA>
<DATA ref="#dynamic.miscdata"
  optional="yes">
  <CATEGORIES><purchase/>
</CATEGORIES>
</DATA>
</DATA-GROUP>
</STATEMENT>

```

Following Algorithm 2, the <STATEMENT> is set as the root or $ESL = 0$. The elements for $ESL = 1$ are found:

$$\begin{aligned}
 ES_1(ESL = 1) &= (purpose, recipient, retention, data-group). \\
 ES_2(ESL = 1) &= (purpose, recipient, retention, data-group).
 \end{aligned}$$

These elements are then reached one by one. As the first element in <PURPOSE> for both statements, the children are checked and matching elements are compared with Algorithm 1. As a result, <current/>, <admin/> in both statements are equivalent and therefore they stay. The third element <contact/> is optional, hence Algorithm 3 is invoked to resolve it and the output is <contact required = "opt-out"/> as Ovalue for opt-out is 3, which is greater than the Ovalue of opt-in (Ovalue = 2). After all children of <PURPOSE> are reached, the

second element <RECIPIENT> at $ESL = 1$ is checked. for ESL greater than 2, we invoke Algorithm 3 for recursive calls. This will continue until the last element is reached. Consequently, we obtain the aggregated privacy policy:

$$Pol(statement) = Pol(purpose, recipient, retention, data-group)$$

V. CONCLUSION

We introduced a new notion of privacy policy aggregation for P3P, which has not been previously explored. We presented a framework for handling P3P privacy policies for service aggregation, which is seen as an emerging technology for providing efficiency and quality of web services. We formally defined the syntax and semantics of our privacy policy aggregation language and provided algorithms based on the formal definitions of privacy policy aggregation. We presented an P3P example to demonstrate how our scheme works. We found that our framework captures all necessary needs for privacy policy aggregation.

REFERENCES

- [1] R. Kanneganti and P. Chodavarapu, "SOA security," in *Proceedings of SACMAT'07*. Manning Publications Co. Greenwich, CT, 2008.
- [2] R. R. Khalaf and F. Leymann, "On web services aggregation," in *Technologies for E-Services 2003, LNCS*. Springer, Heidelberg, 2003, pp. 1–13.

```

<STATEMENT>
  <PURPOSE>
    <current/> <!-- ai = bj -->
    <admin/> <!-- ai = bj -->
    <contact required = "opt-out"/>
    <!-- ai(<contact required="opt-out">) > bj("opt-in") -->
  </PURPOSE>
  <RECIPIENT>
    <ours/> <!-- ai = bj -->
    <delivery/> <!-- ai = bj -->
  </RECIPIENT>
  <RETENTION>
    <indefinitely/>
    <!-- bj(<indefinitely/>) > ai(<stated-purpose/>) -->
  </RETENTION>
  <DATA-GROUP>
    <DATA ref="#user.name"/> <!-- ai = bj -->
    <DATA ref="#user.home-info.postal"/>
    <!-- aj(<"#user.home-info.postal">)>bi(empty) -->
    <DATA ref="#dynamic.miscdata optional="yes">
    <!-- bj(optional="yes" > ai(empty) -->
      <CATEGORIES>
        <content/>
        <!-- bj(<content/>) > ai(<online/>) -->
      </CATEGORIES>
    </DATA>
    <DATA ref="#dynamic.miscdata">
      <CATEGORIES>
        <purchase/>
        <!-- bj(<purchase/>) > ai(<empty/>) -->
      </CATEGORIES>
    </DATA>
    <DATA ref="#dynamic.miscdata">
      <CATEGORIES>
        <preference/>
        <!-- ai(<preference/>) > bj(<empty/>) -->
      </CATEGORIES>
    </DATA>
  </DATA-GROUP>
</STATEMENT>

```

Figure 4. Example of the aggregated policy.

[3] A. I. Anton, E. Bertino, N. Li, and T. Yu, "A roadmap for comprehensive online privacy policy management," *Communications of the ACM*, vol. 50, pp. 109–116, 2007.

[4] E. C. Lupu and M. Sloman, "Conflicts in policy-based distributed systems management," *IEEE Transactions on Software Engineering*, vol. 25, pp. 852–869, 1999.

[5] G. Karjoth and M. Schunter, "A privacy policy model for enterprises," in *Proceedings of the 15th IEEE CSFW'02*. IEEE, 2002, pp. 271–274.

[6] B. Berendt, S. Preibusch, and M. Teltzrow, "A privacy-protecting business analytics service for online transactions," *International Journal of Electronic Commerce*, vol. 12, pp. 109–116, 2008.

[7] T. Yu, N. Li, and A. I. Anton, "A formal semantics for P3P," in *Proceedings of ACM Workshop on Secure Web Services*. ACM, 2004, pp. 1–8.

[8] R. Agrawal, J. Kiernan, R. Srikant, and Y. Xu, "An xpath-based preference language for p3p," in *Proceedings of WWW'03 2003*. ACM Press, 2003, pp. 629–639.

[9] P. Mazzoleni, E. Bertino, and B. Crispo, "XACML policy integration algorithms," in *Proceedings of SACMAT'06*, 2006, pp. 219–227.

[10] "A P3P Preference Exchange Language 1.0 (APPEL1.0)," <http://www.w3.org/TR/P3P-preferences/>.

[11] M. Backes, W. B. G. Karjoth, and M. Schunter, "Efficient comparison of enterprise privacy policies," in *Proceedings of SAC'04*. ACM Press, 2004, pp. 375–382.

[12] A. H. Anderson, "A comparison of two privacy policy languages: EPAL and XACML," in *Proceedings of SWS'06*, 2006, pp. 53–60.

[13] Y. H. Li, H.-Y. Paik, and B. Benatallah, "Formal consistency verification between BPEL process and privacy policy," in *Proceedings of PST 2006*, 2006, pp. 1–10.

[14] M. Backers, M. Durmith, and R. Steinwandt, "An algebra for composing enterprise privacy policies," in *Proceedings of ESORICS 2004*. LNCS 3193, 2004, pp. 33–52.

[15] S. Gevers and B. D. Decker, "Privacy friendly information disclosure," in *Proceedings of OTM Workshops 2006*. LNCS 4277, 2006, pp. 636–646.

- [16] P. Bodorik, D. Jutla, and M. X. Wang, "Consistent Privacy Preferences (CPP): Model, semantics, and properties," in *Proceedings of SAC'08*, 2008, pp. 2368–2375.
- [17] D. Lin, P. Rao, and E. Bertino, "An approach to evaluate policy similarity," in *Proceedings of SACMAT'07s*, 2007, pp. 1–10.
- [18] "The Platform for Privacy Preferences 1.0 (P3P1.0) Specification, W3C Recommendation (April 2002)," <http://www.w3.org/TR/P3P/>.
- [19] "OASIS. security services technical committee. extendible access control markup language (XACML) version 2.0," 2006, <http://docs.oasis-open.org/xacml/xacmlrefs.html>.
- [20] I.V.Ramakrishnan and R. S. W. Xu, "On supporting active user feedback in P3P," in *Proceedings of 2nd Workshop on Secure Knowledge Management (SKM '06)*, 2008, pp. 1–6.
- [21] H. Hochheiser, "The platform for privacy preference as a social protocol: An examination within the U.S. policy context," *ACM Transactions on Internet Technology*, vol. 2, pp. 276–306, 2002.
- [22] L. F. Cranor, *Web Privacy with P3P*. O'Reilly & Associate Inc, 2002.

Dynamic Music Lessons on a Collaborative Score Annotation Platform

Véronique Sébastien, Didier Sébastien, Noël Conruyt
 IREMIA - Laboratoire d'Informatique et de Mathématiques, EA2525
 University of Reunion Island
 Saint-Denis, Reunion (FRANCE)
 veronique.sebastien/didier.sebastien/noel.conruyt@univ-reunion.fr

Abstract - The recent progress in Information and Communication Technologies gave birth to advanced applications in the field of instrumental e-learning. However, most of these applications only propose a limited number of lessons on predetermined pieces, according to the vision of a single music expert. Thus, this article introduces a web platform to create music lessons dynamically and collaboratively, with the assistance of a semi-automatic score annotation module: @-MUSE. To do so, we first describe a new methodology to design such a platform: Sign Management. Then, we detail its general architecture as an Iterative Sign Base System based on a common practice in music learning: score annotation. Lastly, we give various algorithms to generate relevant annotations (explanations) on a score, based on the analysis of musical patterns difficulty.

Keywords - *e-learning; music; knowledge management; sign management; multimedia; annotation; semantic web; ontology; digital score; piano; human-computer interaction; logic; inference*

I. INTRODUCTION

Information and Communication Technology for Education (ICTE) expanded rapidly these last years. Indeed more and more teachers resort to platforms such as Moodle or Blackboard to design their own online courses. While this trend is being confirmed in academic subjects such as mathematics and languages [7], it remains rare for know-how transmission and sharing, for instance in the field of music learning. Indeed, know-how transmission requires heavy multimedia usage and interaction to show the “correct gesture” and is thus complex to implement.

Some instrumental e-learning solutions exist in the form of offline tools, such as instructional DVDs (see the technical report of E-guitare [16]), or business software (Guitar Pro [17], Garage Band [18]). Nevertheless, getting a feedback is capital in know-how acquisition (is my gesture or fingering correct ?). But few applications try to implement a learner to teacher communication axis through video upload and commentaries on the web (see the FIGS [19] glosses system).

Still, the lessons provided by these platforms remain limited to a fixed list of pieces. Although a student can suggest a new title, the realization of a whole lesson on these platforms requires heavy installations and treatments (multi-angle video recording, 3D motion capture), as well as the intervention of multiple actors other than the teacher

himself. While these methods produce high quality teaching material, the realization of a new course remains a complex and expensive process. In parallel, several teachers, for instance retired experts, wish to transmit their know-how in a simple way, without any constraint on the recording location and time and with minimal tool appropriation.

We thus introduce in this paper a complementary framework to rapidly create dynamic music lessons on new pieces with the assistance of a score annotation module.

This framework is implemented on a collaborative score annotation platform for music learning called @-MUSE (Annotation platform for MUSical Education). As described in [11], an online annotation system is chosen because it allows musicians to work with digital scores in a way similar to traditional lessons, where scores are a support for memory and information sharing. In addition, the digital transposition of this common practice enables to enrich it with multimedia incrustation, collaborative working and mobility. As such, its aim is also to constitute a scalable music playing knowledge base to collect and share tips and performances on all possible artistic works referenced on music data warehouse such as MusicBrainz.org [20], and which can evolve according to the learners’ needs. This base is called ISBS (Iterative Sign Base System).

In this paper, we first introduce the methodology and principles of Sign Management that supports this platform. Then, we describe the general architecture of @-MUSE, based on Semantic Web concepts, in order to constitute a musical sign base (ISBS). To assist users into feeding and exploiting this base, we describe various methods to generate relevant annotations (i.e., explanations) on a score. Lastly, we conclude this work by detailing its principal perspectives: an adapted tactile interface and some serious gaming aspects.

II. METHODOLOGY : SIGN MANAGEMENT

Sign Management deals with the management of know-how rather than knowledge. It manages live knowledge, i.e., subjective objects found in interpretations of real subjects on the scene (live performances) rather than objective entities found in publications (bookish knowledge). A Sign is a semiotic and dynamic object issued from a Subject and composed of three parts, Data, Information and Knowledge. All these subjective components communicate together to build a chain of sign-ifications that we want to capture.

Sign management is thus more central than Knowledge management for our purpose in instrumental music learning. Indeed, the musical signs to treat are made of emotional content (performances), technical symbols (scores) and tacit knowledge (rational and cultural know-how). Thus, a Sign is the interpretation of an object by a subject at a given time and place, composed of a form (Information), a content (Data) and a sense (Knowledge). The sign management process that we have created is made on a Creativity Platform for delivering an instrumental e-learning service [10][4][5]. It is founded on an imitation and explanation process for understanding gestures that produce a right and beautiful sound. The advantage for learners is that we are able to decompose the teacher’s movement and understand the instructions that are behind the process of playing a piece of music. In fact, a lovely interpretation is made of a lot of technical and motivated details that the learner has to master, and the way we want to deliver this information is to show examples from experts through multimedia annotations indexed on the score. To do so, we introduce a new platform to design dynamic music lessons through multimedia annotations: @-MUSE.

III. @-MUSE GLOBAL ARCHITECTURE

As the aim of @-MUSE is to enable dynamic teaching and learning, it is capital that its architecture remains flexible. The usage of Semantic Web tools is thus an appropriate lead to allow the platform to benefit from a “networking effect”. Indeed, a significant amount of scattered musical resources already exist on the web and can be relevant in the context

of music lessons. These resources can be music metadata (MusicBrainz.org), digital scores (images, PDFs, MusicXML free or proprietary files available on Werner Icking Archive [21]), multimedia documents (recordings of video performances and lessons on YouTube [22] or eHow [23]) or simple textual comments. They constitute the different sign components listed in part II: data, information and knowledge. As many of these resources benefit from a Creative Commons License [24], they can be used in the context of a music lesson, complementary to high quality resources from a professional multimedia capture set [5]. Figure 1 exposes a comparison between traditional instrumental e-learning applications architecture and @-MUSE architecture. In the first case, lessons are defined in a static way. Each lesson correspond to a musical piece, with its associated resources : video, audio and image files synchronized together to form the lesson. While this system produces complete lessons, it cannot establish relations between two distinct resources or pieces, which is an essential point when learning music as a whole. In the second case, @-MUSE dynamically creates lessons by linking related resources and presenting them to the user in an adapted interface [11]. If a resource is not available (for instance, a logic representation of a score), the system still works with a temporary replacement (for instance a simple image representing the score) in the frame of a degraded mode. It can then point to any user the need to provide such resource to enable new functionalities on the platform. As more links are created between resources, different representations of the same piece can be proposed to learn

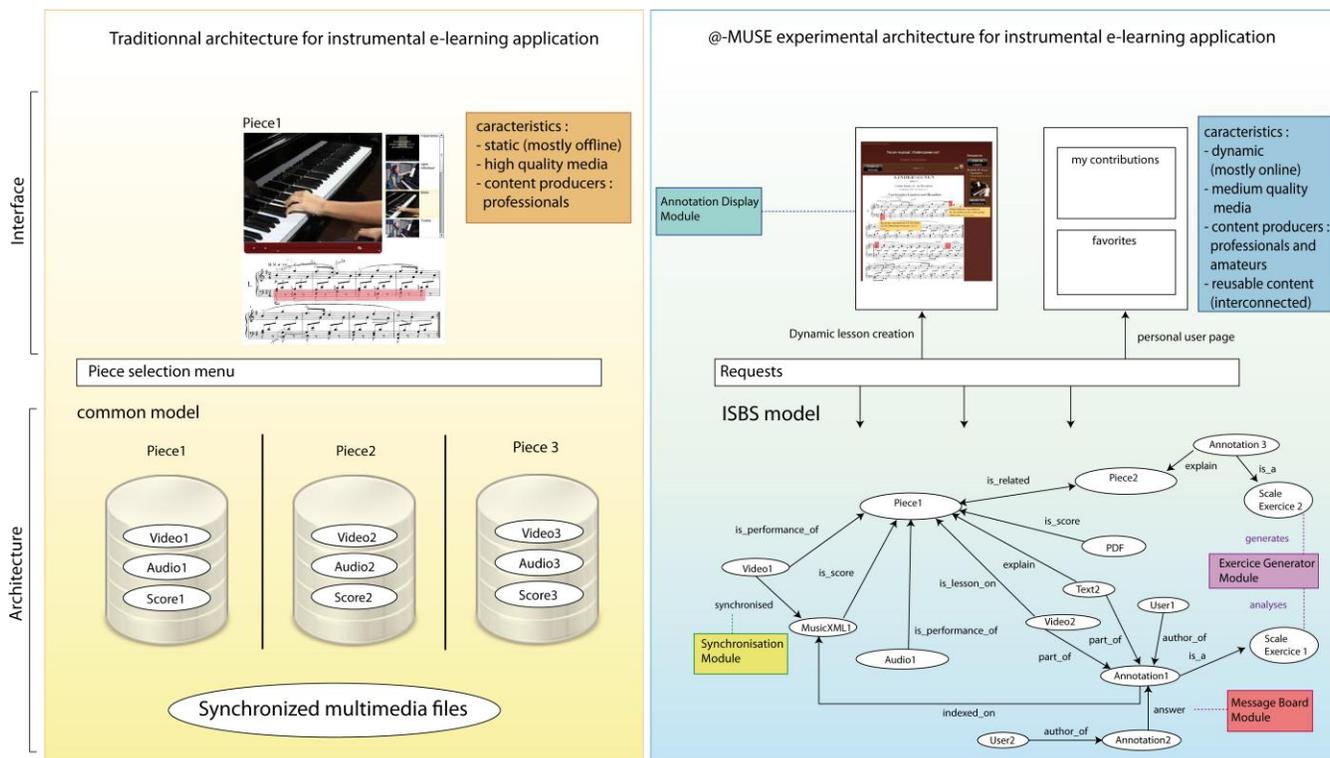


Figure 1. Architecture comparison between traditional instrumental e-learning application and @-MUSE

how to play it. Some links such as a time synchronization between two representations (i.e., a video performance and a logical description of the score) can be realized by specific independent modules (see Figure 1).

We have done previous work in [12] to propose an adapted ontology to link musical resources in an educational context using the Resource Description Framework (RDF [8]).

In the end, the association of these elements will allow the creation of an Iterative Sign Base System in the same vein as IKBS (Iterative Knowledge Base System [3]). The difference here lies in the manipulation of semiotic objects (signs), instead of conceptual ones (knowledge), as described in part II. The following chapter explains how new signs can be generated on this platform through semi-automatic score annotation, and thus participate in the enrichment of the sign base (ISBS) by demanding minimal efforts from the platform users.

IV. INFERENCE ON DIGITAL SCORES

ISBS is a sign base model designed to collect musical signs such as scores (model) and performances (cases), in order to explain and compare them. To realize such analysis in a semi-automatic way, we need to detect specific patterns within a score. This detection could be made directly on performances [14] but audio signal analysis algorithms are difficult to implement in a web application and may be unreliable in an educational context. That is why we rely on XML representations of a score. MusicXML [1] is an XML open source format to describe digital scores staff by staff, measure by measure, and lastly note by note (Figure 2).

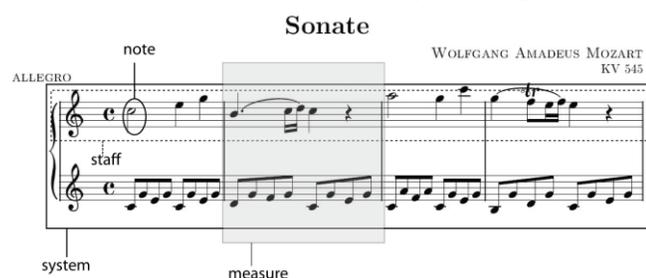


Figure 2. Score logical structure

In what follows, we review and propose different methods to extract various playing information from a piece metadata and structure.

We base these methods on how a pianist would address an unknown piece. As detailed in the descriptive model presented in [12], the musical work is first replaced in its context (composer, period, form). Then, its difficulty is evaluated, firstly globally, and then part by part, in order to determine what type of work can be made on this piece and where.

Thus, the first playing related information we display on a new piece is an approximation of its difficulty. In TABLE 1, we propose seven criteria affecting the level of a piece for piano and detail how they can be estimated from a MusicXML file. Globally, a piece difficulty depends on its tempo, its fingering, its required hand displacements, as well

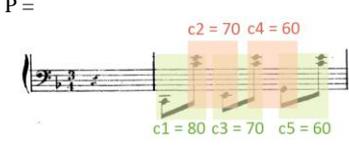
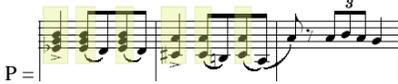
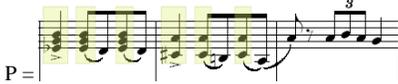
as its harmonic, rhythmic and polyphonic specificities. Of course, these various criteria affect each other in a complex manner. For example, hand displacement is strongly affected by fingering, as noted in TABLE 1.

Indeed, among these seven criteria, fingering plays an important role. Several works present methods to automatically deduce fingering on a given musical extract for piano ([2][9][6]). Most of them are based on dynamic programming. All possible fingers combinations are generated and evaluated, thanks to cost functions. The latter are determined by kinematic considerations. Some functions, like in [6], even consider the player's hand size to adjust its results. Then, expensive (in term of effort) combinations are suppressed until only one remains, which will be displayed as the resulting fingering. While the result often differs from a fingering determined by a human professional, it remains largely playable and exploitable in the frame of an educational usage. However, few algorithms can process polyphonic extracts [6], and many other cases are ignored (i.e., left hand, finger substitutions, black and white keys alternation).

Even if more work is needed on this issue, the use of cost functions remain relevant as it is close from the process humans implicitly apply while working on a musical piece. That is why we extend this idea and create complementary criteria to design a piece difficulty analyzer for piano learning. For each criterion described in TABLE 1, a score is calculated in percentage. The piece difficulty rate is thus the average rate of each criterion. Furthermore, some weighting coefficients can be affected to each criterion to reflect the particularities of the player. For instance, pianists who are really at ease with polyrhythm would not consider it a relevant factor, thus affecting it a 10% weight.

However, we insist that the resulting difficulty rate should be interpreted with care and remains a simple approximation. As stated in [15], a pleasant performance is not a mere addition of criteria since it contains an important subjective part. Moreover, for the time being, the algorithms we propose remain bold and need some specific refinements which will be the object of a next paper. Indeed, some cost functions are applied measure by measure, while they should be applied phrase by phrase to remain coherent with the piece logic. Also, some of the parameters were determined after the practices of a small group of advanced pianists and need to be extended by working with a larger sample of musicians, including other instruments.

TABLE 1. PLAYING DIFFICULTY CRITERIA IN PIANO PRACTICE

Performance difficulty criterion	Musicological definitions	Cost function definition	Examples	MusicXML implementation
Playing speed	Tempo: speed or pace of a musical piece. May be indicated by a word (ex: allegro) or by a value in BPM (Beats Per Minute) Pulsation: reference value indicated in the tempo : $\text{♩} = 1$, $\text{♪} = 2$, $\text{♫} = 4$, $\text{♬} = 8$, $\text{♭} = 16$, etc.	Playing speed = tempo / (shortest note value in the piece) Unit: beats (time value)	P1: tempo = 120 ♩ Shortest value = ♩ P1 playing speed = $120 * 8 / 16 = 60$ P2: tempo = 120 ♩ Shortest value = ♩ P2 playing speed = $120 * 4 / 16 = 30$ Conclusion: Some parts in P2 are played faster than in P1. To be more accurate, it is possible to multiply the result by the proportion of notes of shortest value. Thus, if P1 contains 40% ♩ , and P2 only 5%, then P1 is globally faster.	<note><type> elements Tempo attribute in <sound> element
Fingering	Fingering: choice of finger and hand position on various instruments. Different notations exist according to the instrument. (Ex: in piano: 1 = thumb, 2 = index finger, 3 = middle finger, etc.)	If m_1, m_2, \dots, m_n represent the measures of a given piece P, $\text{Fingering_difficulty}(P) = \sum (\text{Fingering_cost}(m_i) > 50)$ See [2][6][9][13] for more detail.	 P =  Fingering_cost(m_1) = 10 Fingering_cost(m_2) = 0 Fingering_cost(m_3) = 70 Fingering_difficulty(P) = 70	<measure> and <note> elements
Hand Displacement	Interval: pitch distance between two notes, in semitones. A hand displacement is considered difficult when two successive notes (or two chords) are spaced by at least 7 semitones, played by close fingers (on the same hand, distance < 4 fingers) at a high tempo. The displacement cost of an interval increases with its gap length. It also increases with polyphony.	If d_1, d_2, \dots, d_n represent n intervals verifying the conditions given in the previous description, in a piece P $\text{Displacement_difficulty}(P) = \sum \text{Displacement_cost}(d_i)$	P =  Displacement_difficulty(P) = 340	Combined <note> elements where <pitch> gap ≥ 7 . Associated fingering file.
Polyphony	Chord: aggregate of musical pitches sounded simultaneously.	Proportion of chords and chords sequences in the piece	 P =  Chords_proportion(P) = $6 / 16 = 38\%$	<chord> element
Harmony	Tonality: system of music in which specific hierarchical pitch relationships are based on a key "center", or tonic. Various tonalities impose various sharps and flats as a key signature. The most basic ones (no alteration) are A minor and C major.	Proportion of altered notes	 P =  Altered_notes_proportion(P) = $3 / 25 = 12\%$	<alter> and <accidental> elements
Irregular Rhythm	Polyrhythm: simultaneous sounding of two or more independent rhythms. Example : synchronizing a triplets over duplets	Proportion of remarkable polyrhythm patterns (Time reference = pulsation)	 P =  Polyrhythm_proportion(P) = $4 / 4 = 100\%$	<time-modification> element
Length	The length of the piece in beats. NB: the number of pages cannot really reflect the length of a piece because of page setting parameters	Number of measures * number of beats per measure.	 P =  Length(P) = $3 * 3 = 9$	<beats> element of <time> element and <measure> elements

This algorithm also serves as a base for the next chain of information inference on the given piece. Indeed, it can be applied to identify difficult parts within the piece. By calculating the difficulty rate of each measure, we can display the remarkable parts, which rate exceeds a given threshold (determined by the player's level). The cause of its difficulty can then be deduced from the rates of each criterion (Figure 3). The application can then annotate the part accordingly, for instance by redirecting the learner to an adapted exercise.

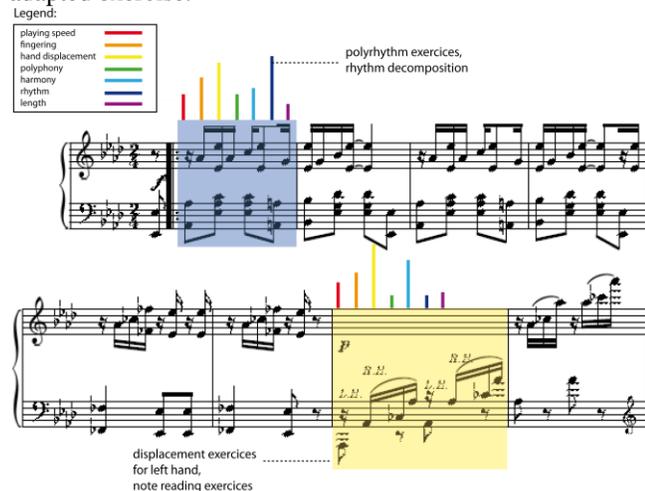
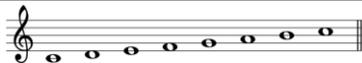


Figure 3. Difficulty analysis and recommendations on a digital score

In parallel to difficult parts, other remarkable structures can be identified within a piece. Indeed music learning relies a lot on the repetition of specific short patterns, with slight differences (for example the tone of a piece), which can be reused in various context, especially within the same genre (baroque, classical, jazz, etc). TABLE 2 gives some patterns examples.

TABLE 2. MUSICAL PATTERNS EXAMPLES

Pattern name	Example
Scale	
Arpeggio	
Trill	
Real sequence	

If long enough (and thus actually remarkable), each of these patterns can be detected as a note sequence within a MusicXML file. Then, corresponding exercises can be pointed to guide the learner. These exercises can be directly adapted from the considered pattern. For instance, in the case of an arpeggio, the latter will be extended to the whole keyboard and repeated part by part, by adding a new note every ten repetition. This process can easily be computed as suggested by Figure 4. Anytime, the annotation's owner and teachers can modify it in order to improve the given explanation with textual and video commentaries, symbols and tags. Users can also invalidate the generated annotation if considered as inappropriate. In this case, the motive for the suppression should be specified. This data will be later used to determine the reasoning error in order to improve the next generated annotations. This process will be detailed in an upcoming paper.

V. CONCLUSION AND PERSPECTIVES

In this paper, we proposed a methodology (Sign Management), a model (Iterative Sign Base System) and some inference methods to build an instrumental e-learning platform called @-MUSE. This platform allows teachers and learners to create music lessons dynamically with the assistance of a semi-automatic pieces annotator. These lessons can evolve according to the users' needs by submitting contextual exercises to them, in the form of multimedia annotations. These exercises are generated from the original score based on the identification of remarkable patterns and their playability. Users can then give their point of view on the generated annotations but also add new ones, thanks to a dedicated symbols library as well as a multimedia capture module. The more knowledge is created on the platform, the more detailed will be the lessons, thanks to the emerging network effect resulting from the semantic linking of the various resources.

Different perspectives are also considered for this work, including the addition of tactile functionalities, as well as some serious gaming aspects. For instance, an interface adapted to tablet PC would allow to use our platform directly in front of the instrument, guaranteeing an experience close to a traditional music lesson. The collaborative aspects of such a platform also need to be studied to approach music learning under an entertaining angle, for instance by proposing specific group performances (Global Sessions [25]) and game features. Indeed, as implied by our platform's name, learning music should first and foremost be a pleasure.

Towards a Unified User Profiling Scheme for Distributed Large Sporting Events' Environments

Uko Asangansi, Stefan Poslad
 School of Electronic Engineering and Computer Science,
 Queen Mary University of London
 London, United Kingdom
 {uko.asangansi, stefan}@eecs.qmul.ac.uk

Abstract— This paper presents a user modeling and personalisation framework for providing personalized services to users through their mobile devices during large sports events. The user model combines the knowledge of sports events the user physically attends and the knowledge of the user's interaction behavior when consuming multimedia content from his mobile while away from the sporting event's venue(s). The user model employs both explicit and implicit modeling techniques which is able to learn and represent shifts in the user's preferences. Ontologies are used to formalize the user model and domain knowledge thereby disabling ambiguities in preferences specification but introducing reasoning capabilities.

Keywords- Ontology, user profile, sports, olympics

I. INTRODUCTION

Distributed large sports events (DLSE) like the Summer Olympics, Commonwealth Games, and Paralympics are often characterized by several sport events of diverse disciplines taking place at disparate venues and places and spanning a number of days. This high activity environment is becoming a melting pot for several dimensions of research into personalization in order to enable spectators make better use of the enormous amount of information that is often generated in this sort of environment [7]. Mobile devices lend themselves as a target platform for personalisation in this environment due to their small form, light weight [3] and sensors, e.g., GPS (Global Positioning System) modules which can be queried to return a device's position.

Current state of the art of personalisation systems in DLSEs can be broadly classified into two categories based on their target context of usage. While the first category focuses on the spectator who physically attends the sports events, the second category focuses on remote users who receive the content over a network. On the former category, the personalisation system delivers complementary multimedia information services, e.g., replays, game statistics etc. but these systems do not factor in other sports events the user might be interested in and has accessed through their devices. In the second category, emphasis is laid on the user's interaction with the networked (mobile) device and viewing pattern in order to deliver a personalized service but no consideration is given to what events the user has physically attended in the past. In this paper, we present a user modeling and personalisation framework that attempts to unify both worlds as they combine the sports multimedia

consumption habit of users on their mobile devices and information on what sporting events they physically attend during DLSEs inferred through the user's location. This combination is done in order to offer the user a more accurate personalisation service within and without the sports arena.

In order to investigate this, we propose an ontology-based location augmented user profiling scheme which profiles users attending sports event by monitoring the sporting events content they consume within and without the events they attend. An ontology based approach is chosen because it provides a richer representation scheme which is more precise and less ambiguous than other conventional schemes [6]. In addition, it provides an adequate foundation for the representation of coarse user interests to fine-grained preferences in a hierarchical way, and can be a key enabler to deal with the subtleties of user preferences [4].

In the rest of this paper, a review of related work is presented in Section II and it is followed by the architecture proposed in Section III. The constituent ontologies of the system and the applied proposed profile learning techniques are specified in Section IV and V respectively and finally, some future work is discussed.

II. RELATED WORK

There has been several research endeavors directed towards enriching users' interest, engagement, and experience of DLSEs in different ways. In order to deliver this experience, these systems need to maintain an internal model of the user.

In [12], a personalized live sports event viewing system for mobile devices is presented. The system uses an implicit user model driven approach to enable the personalisation system which adaptively predicts a user's preferred events during live sports shows. While this work demonstrates how a personalisation system can use a user's previous viewing session to elicit the user's preferences, it does not take the user's location (which can offer more insight into the user's preferences) into consideration. In addition, a simple XML format is used to represent the user model. Hence not much of reasoning can be done with such a model.

In [2], a place-shifted sports 'snacking' application is presented. The system provides sports fans a medium to catch-up with their favorite sports when they cannot view the live event in person as it unfolds. Although this system does not explicitly describe its user model, one can envisage that a very simple user model is used to achieve this.

In [1], a mobile based in-stadium information system is presented. The system provides on-demand instant sports replay and traditional media convergence functionality to mobile devices by capturing and processing a television signal. This system assumes a generic user model for all users and does not support any out-of-stadium experience. Other DLSE personalization systems include [10], which allows users to follow athletes of their choice by tracking their preferred athlete’s location, speed and pulse. This idea is pushed further by [5], where users can track their preferred cyclists in a long distance cycling race and based on the tracking information, they can decide where to position themselves in order to get a good view of their preferred athlete. Although these works use the concept of location, it is used in improving the user’s interaction with the associated sport and does not in any way aid the personalisation system in gaining a better knowledge of the user’s preferences.

The aforementioned works show the lack of a rich user profiling scheme in the prevailing DLSE personalization domain. In order to develop a functional user profiling scheme, some requirements as outlined in [8] ought to be considered. Some of these requirements include:

Semantic Reasoning – the user model and multimedia artifacts (video, text, pictures, metadata etc) should be semantically modeled based on ontologies in order to enable semantic reasoning.

Dynamic User Profiling - human preferences generally tend to be dynamic and transient. Most of this dynamic behavior can be captured by continuously monitoring and recording the user’s behavior for analysis. This record can be collected in form of a history of multimedia artifacts assessed or requested by the user as a function of the time spent consuming it. In addition this work argues that a location component ought to be added to this record.

III. SYSTEM ARCHITECTURE

The aim of this proposed framework is to formally elicit a location-augmented user model optimized for mobile devices aimed at large spots events. The architectural overview of the proposed framework as depicted in Figure 1 shows the main components of the system and their respective data flow. As shown, the proposed system architecture consists of a User Profile Management System, a Personalised Event Listing Service (PELS), a Sport Events Schedule Server and some domain ontologies. A brief outline on these components is as follows:

The User Profiling system consists of the User Profile Ontology, Profile Learning Module, the (Runtime) User Profile, the Profile Store and the Profiling Proxy – which runs on the personalized application installed on the user terminal. These components work together to ensure that both explicitly and implicitly collated information from the user is used to identify and represent the user’s static and transient interests or preferences. The PELS is responsible for matching users with available live sport events as described in the semantically annotated event schedule and it

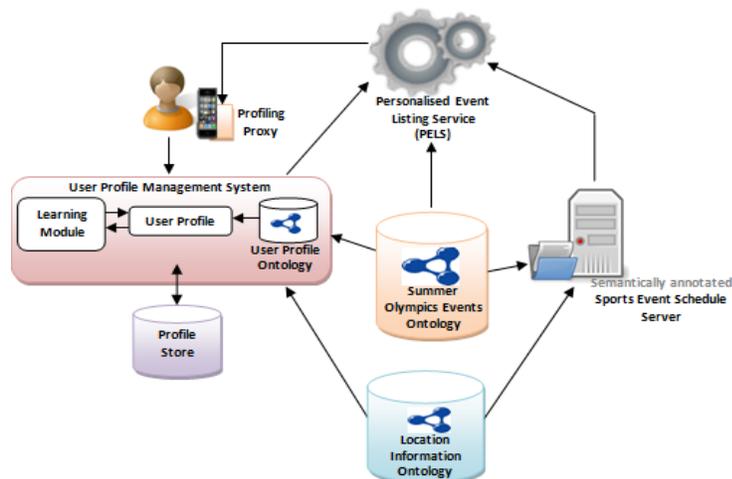


Figure 1. Overview of the proposed framework

outputs a ranked list of live events that matches the user preferences as specified in the user’s profile. This is a ranked list ordered by the degree of relevance it has to the user. The associated domain Ontologies are integrated for knowledge inference and vocabulary control. In line with this, users can only express their preferences on concepts formalized in the Summer Olympic Events (SEO) ontology and concepts from this ontology are used in for annotating the schedule.

A. The Personalization Process

In a nutshell, the personalisation process as offered by the PELS begins when the user logs-in into the system. A successful log-in sub-process involves the transfer of a profiling proxy from the user’s mobile device to the profiling service. The profiling proxy at log-in encapsulates the user’s unique identifier, current location, usage history (of the last session) and location history etc. The proxy remotely and temporarily stores the user’s usage information before it is sent to the Profiling system – this is done periodically. The information from the proxy is used to uniquely identify the user and retrieve his full profile from the profile store. The retrieved profile is then updated with the content of the profiling proxy (depending on which of the information is more current). The updated profile is then passed to the PELS.

The PELS receives the user’s profile and reasons with it along with a semantically annotated schedule of the live sport events which includes the event venue for every given event. The output of this reasoning process is a ranked list of events that matches the user’s interests and relevant to the user’s current location.

IV. DOMAIN ONTOLOGIES

The domain ontologies specified in the proposed framework are responsible for giving structure to the knowledge represented in the framework. The ontologies conceived as depicted in Figure 1 include the Summer Olympic Events (SEO) ontology, the Location Information Ontology and the User Profile Ontology. In developing these ontologies, the Web Ontology Language (OWL) was used to

formalize an ontology conceptualization and produce a hierarchy of concepts. The concepts and relationships for the Summer Olympics Events domain ontology in this framework are specified based upon the structure of sports information collated from the Olympics organization website and the rules and regulation handbook for each respective sports discipline – retrieved from the website of the respective governing body for the given sport. While the Location Information Ontology contains concepts and relations that represent location information but only focusing on those that can be used to connect the domain knowledge (which include the sport events and their respective venues) with user’s location(s) during the profile learning and personalisation process. On the other hand, the User Profile Ontology comprises concepts and relations that represent a user, his preferences and past behaviors.

A. The Summer Olympic Events (SOE) Ontology

The Summer Olympics Events Ontology represents the conceptualization of the domain of Summer Olympics Events. The aim of this ontology is to provide a vocabulary and background knowledge for eliciting user preferences as well as annotating the events schedule in order to streamline the semantic matching and personalisation processes in the PELS. Given that users are currently only allowed to specify their preferences for sport events and venues, athletes and sports officials are currently not included in the model.

The SOE Ontology is modeled to be consistent with the sport events nomenclature and taxonomy used by the International Olympics Committee (IOC) which has an Olympic sport at the top of the taxonomy, then sports’ disciplines and the sports events in the bottom. According to the IOC, a sport is a single or group of disciplines regulated by an international federation (a governing body).

Taking the IOC’s model into consideration, the SEO ontology was conceptualized to include all 26 Olympic Sports, and their respective disciplines, and events in the Summer Olympic Games. The sports are modeled as top level concepts, while the disciplines as the subclasses (where applicable) of these top level concepts while the events are instances of the subclasses. In addition, each individual event is qualified by a number of object properties such as *performedWith*, *performedIn* in order to support a richer inference and classification scheme by comparing the values of the specified object properties.

B. Location Information Ontology (LIO)

The Location Information (LIO) domain ontology conceptualizes the notion of location in the proposed framework. Location related information is necessary for adding a location component to the semantically annotated events schedule thereby signifying where the events are taking place which translates into providing a means to model location information for the events described in the SEO domain ontology. Thereby, enabling the personalization system to filter events according to a user’s current location. Moreover, users can elicit their preferences through locations (venue of sports events) explicitly – by users choosing which sport event venues they will prefer to visit or receive live

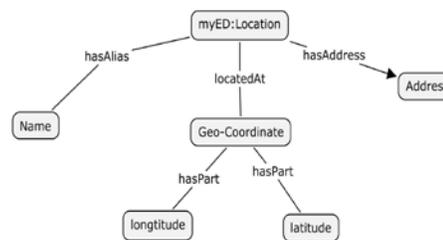


Figure 2. An abridged representation of the Location Information Ontology (LIO).

information from or implicitly – by the system monitoring which event venues the user actually attends and using the information as a form of relevance feedback to augment the user’s general preference model. Figure 2 shows an abridged model of the LIO.

The Location concept is has the property *hasAddress* which qualifies the address of the given location. The *locatedAt* property reflects the location’s geo-coordinates (in logtitude and latitude) while the *hasAlias* is used to hold the location’s well-known name e.g., Aquatics Centre.

C. User Profile Ontology (UPO)

In conceptualizing the User Profile Ontology (UPO), we were inspired by a number of user models especially the General User Model Ontology [11]. However, we have so modeled it such that only concepts relevant to our domain of discourse are retained. Our model logically contains both static and dynamic representations of the User. The static information are those that are less likely to change over time and this is represented by the Person the user is. This sub-model describes the user’s unique identifier and other demographic information. The other components comprise the dynamic entities whose instances are updated by the system as the user interacts with it. These components include the user’s *locationHistory*, *usageHistory*, and Preferences.

Furthermore, as depicted in Figure 3 the proposed User Profile Ontology is modeled to capture not only the sports events selected by the user in previous instances, but also the location where the user was. For instance, the user selecting an Archery event while at the Aquatic Centre may signal the user’s preference for swimming events over archery.

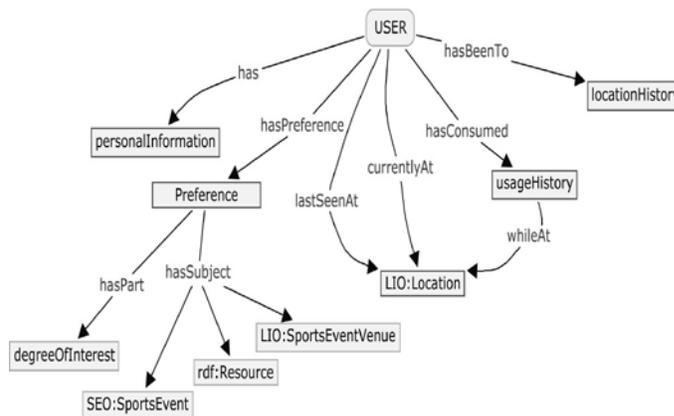


Figure 3 An abridged representation of the User Profile Ontology

The *Preference* concept is used to specify the user's sports event and venue interests and disinterests respectively. These interests are associated with a *degreeOfInterest* index which is used in a preferences weighting scheme and reflects how much the user prefers a given event over the others. This weighing scheme allows the profiling system to dynamically increase the weights associated with sports events the user continually shows an interest in, while the index of those the user shows disinterests in are gradually attenuated.

V. USER PROFILE LEARNING

Since user's preferences change over time [8], it is imperative to for the system to learn the user's preferences by monitoring the user's behavior by analyzing the contents they have viewed in the past as stored in the *usageHistory* component of the profile and the sports venues they have been to as stored in the *locationHistory* component of the profile. One approach employed in ensuring that the user's profile stays relevant is by updating the *degreeOfInterest* component for every preference concept in the user's profile by a time-based decay function based on the user's behavior (i.e., sports events venue attendance habit and mobile application usage). The decay function is as follows:

$$doi_{new} = doi_{old} + Rfb \times e^{-\delta x} \times \log\left(\frac{time}{\log length}\right) \quad (1)$$

The doi_{old} variable stands for the current *degreeOfInterest* component of the preference concept. Rfb is the relevance feedback factor given through an analysis of the content consumption and frequency of visits to a given event venue; the relevance feedback value is taken to be a Boolean value. The $\log\left(\frac{time}{\log length}\right)$ expression reflects the time spent at the given event venue or watching an event content item on the mobile application and the duration of the event or length of the content, operates as the normalizing factor. The $e^{-\delta x}$ factor is used to cushion the personalized non-linear change of the concept's weight according to user *locationHistory* and *usageHistory* data. 'x' represents the number of consumed content. The more content a user consumes, for example, the more slowly the weights increase. The δ factor is a constant, which takes different values in the two opposite scenarios of consumed/non-consumed content. More precisely, in the case of non-consumed content, the changing rate (i.e., the decreasing rate) should be slower, since a non-consumed content does not constitute an explicit indication for non-interest. On the contrary, in case of consumed content the changing rate (i.e., the increasing rate) should be faster, since a consumed content demonstrates a better indication for interest.

VI. FURTHER WORK

In moving this work forward, the profile learning algorithm will be improved by integrating a weight spreading algorithm which will update other semantically related preference concepts' weights when one of the neighboring concepts in the user model is updated. In

addition, an empirical evaluation of the framework will be carried out using real users and the results will be compared with that of convention DLSE personalization systems. Furthermore, aspects of social and group modeling are areas to be further investigated with respect to this work.

ACKNOWLEDGMENT

This work has been undertaken within the framework of the My-e-Director 2012, Real-Time Context-Aware and Personalized Media Streaming Environments for Large Scale Broadcasting Applications, FP7 Project (grant No. 2152482012), funded by the European Commission.

REFERENCES

- [1] A. Ault, J. Krogmeier, S. Dunlop, and E. Coyle, "eStadium: The Mobile Wireless Football Experience," *Internet and Web Applications and Services, International Conference on*, pp. 644-649, 2008 Third International Conference on Internet and Web Applications and Services, 2008 doi:10.1109/ICIW.2008.57
- [2] F. Bentley and M. Groble, "TuVista: meeting the multimedia needs of mobile sports fans," *Proceedings of the seventeen ACM international conference on Multimedia (MM '09)*. ACM, New York, NY, USA, Oct. 2009, pp. 471-480, DOI=10.1145/1631272.1631337
- [3] U. Bhuvan and M. San, "The Enterprise Mobile Applications Development Framework," *IT Professional*, vol. 12, no. 3, pp. 33-39, May/June 2010, doi:10.1109/MITP.2010.45
- [4] D. Vallet, I. Cantador, M. Fernandez, and P. Castells. "A Multi-Purpose Ontology-Based Approach for Personalized Content Filtering and Retrieval," *Proceedings of the First International Workshop on Semantic Media Adaptation and Personalization (SMAP '06)*. IEEE Computer Society, Washington, DC, USA, pp. 19-24.
- [5] A. Devlic, M. Koziuk, and W. Horsman. "Synthesizing Context for a Sports Domain on a Mobile Device," *Proceedings of the 3rd European Conference on Smart Sensing and Context (EuroSSC '08)*, Daniel Roggen, Clemens Lombriser, Gerhard Tröster, Gerd Kortuem, and Paul Havinga (Eds.). Springer-Verlag, Berlin, Heidelberg, pp. 206-219. DOI=10.1007/978-3-540-88793-5_16
- [6] G. Dragan, D. Djurić, and V. Devedzović. *Model Driven Engineering and Ontology Development*. Dordrecht: Springer, 2009. pp.18-23.
- [7] X. Sun and A. May, "Mobile personalization at large sports events user experience and mobile device personalization," *Proceedings of the 2nd international Conference on Usability and internationalization (Beijing, China, July 22 - 27, 2007)*. N. Aykin, Ed. Lecture Notes In Computer Science. Springer-Verlag, Berlin, Heidelberg, pp. 486-495.
- [8] K. Kesorn, Z. Liang, and S. Poslad, "Use of Granularity and Coverage in a User Profile Model to Personalise Visual Content Retrieval," *Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2009. CENTRIC '09. Second International Conference on*, pp.79-84, 20-25 Sept. 2009 doi: 10.1109/CENTRIC.2009.19
- [9] M. Golemati, A. Katifori, C. Vassilakis, G. Lepouras, and C. Halatsis, "Creating an Ontology for the User Profile: Method and Application," *Proceedings of the First RCSIS Conference. RCSIS 2007, Ouarzazate, Morocco, April 2007*.
- [10] D. Olsson and A. Nilsson, "MEP - A Media Event Platform," *Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 9 (HICSS '01)*, Vol. 9. IEEE Computer Society, Washington, DC, USA. pp., 3-6
- [11] D. Heckmann, T. Schwartz, B. Brandherm, M. Schmitz, and M. Wilamowitz-Moellendorff, "GUMO - The General User Model Ontology.," *UM, Edinburgh, UK, July 2005*, pp. 428-432.
- [12] Z. Wang, S. Poslad, C. Patrikakis, and A. Pearmain, "Personalised Live Sports Event Viewing on Mobile Devices," *Ubicomm, Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies, 2009*. pp.59-64

Coordination based Distributed Authorization for Business Processes in Service Oriented Architectures

Sarath Indrakanti Vijay Varadharajan
 Information and Networked Systems Security Research
 Dept. of Computing, Macquarie University, Australia
 {sindraka, vijay}@ics.mq.edu.au

Abstract — Design and management of authorization services in service oriented architectures poses several challenges. In this paper, we propose authorization architecture for business process layer in service oriented architecture. We describe the components and functionalities of the architecture such as authorization policy evaluators, certificate and credential authorities and dynamic attribute services and discuss the security management of these functions at specification time and at run time. Then the paper describes authorization evaluation algorithms and discusses the design choices for evaluation models. Finally, the paper describes the benefits of the proposed architecture, which has been implemented.

Keywords- Authorization, Business Processes, Service Oriented Architectures

I. INTRODUCTION

Broadly speaking, the Service Oriented Architecture (SOA) comprises web services and business workflows built using web services. These workflows are called business processes [1]. Figure 1 shows the positioning of the authorization service components within the various layers of the SOA. Authorization services for the web services layer have special design requirements because web services present a complex layered system. For instance, a service could be a front-end to an enterprise system where the enterprise system accesses information stored in databases and files. Web services may be used by enterprises to expose the functionality of legacy applications to users in a heterogeneous environment. Alternatively, new business applications could be written to leverage benefits offered by web services. This means that authorization architecture for web services must support multiple models of access control. This enables legacy applications to use the access control models they have already been using as well as new web services applications to use new models of access control.

Currently, there exist a range of authorization models for stand-alone systems and traditional distributed systems. There also exist a few authorization schemes that are designed either for the web services layer [2,3] or the business process layer [4, 5] of the SOA. There is no unified model currently available that provides a comprehensive authorization framework for both web services and business

processes comprising the SOA. After carrying out a thorough survey and analysis of the existing authorization models built for stand-alone systems and traditional distributed systems as well as for various layers of the SOA, we have formulated the design requirements for authorization services required for web service and business process layers. These are described in Section II.

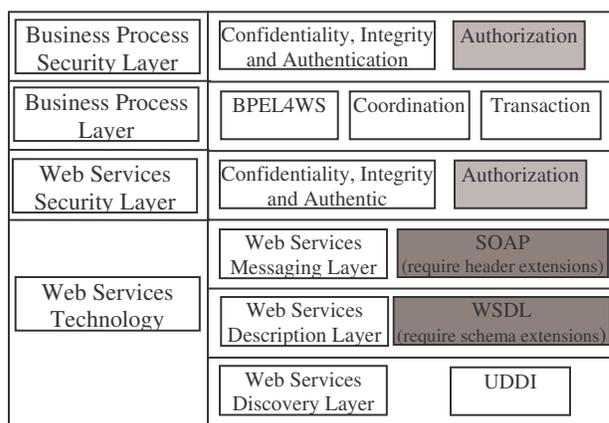


Fig. 1: Layers in the SOA

Taking into account these design principles, we have proposed a unified authorization framework for the SOA. The authorization framework comprises two separate authorization architectures (indicated by the light-grey coloured boxes in Figure 1 that extend the security layers of web services and business processes. Extensions to the web services description and messaging layers are also proposed to support the unified authorization framework for the SOA (indicated by the dark-grey coloured boxes in Figure 1). This work builds on the work on the authorization architecture for web services, referred to as the Web Services Authorization Architecture (WSAA) [6]. In this paper, our focus is on the business process layer authorization architecture, which is referred to as the Business Process Authorization Architecture (BPAA). The BPAA builds on top of the WSAA and forms part of the overall authorization framework for Service Oriented Architectures. BPAA is the focus of this paper.

Authorization architecture for the business process layer of the SOA must provide orchestration services to coordinate the authorization decisions from individual partner's authorization policy evaluators. Each partner must be

allowed to control its own authorization policies and also not require disclosing them to the entire workflow or to the workflow engine. Even in cases where the binding to actual end-points of partner services happens dynamically at runtime, the authorization architecture must be able to orchestrate the partners' authorization policy evaluators and arrive at an authorization decision.

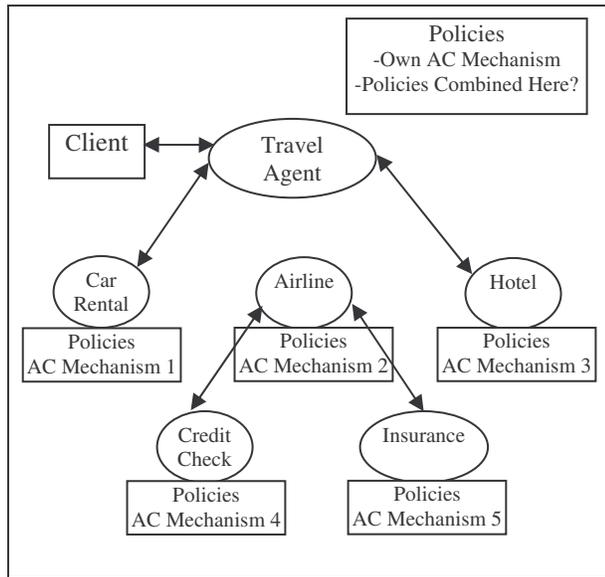


Fig. 2: Travel Agent Service Example

Consider for instance the Travel Agent Service shown in Figure 2, where each of the partner services may potentially use their own access control (AC) mechanisms. The partner airline (for example, United Airlines) may not wish to disclose its policies to the travel agent. Similarly, other partners also may not wish to disclose their policies to be combined by the travel agent in order to authorize the client. In the course of the workflow, the client needs to get authorized seamlessly to partner services.

The paper is organized as follows. In Section II, we consider the requirements for authorization framework for Service Oriented Architectures. Section III describes the proposed business process authorization architecture. Finally Section IV concludes.

II. AUTHORIZATION DESIGN REQUIREMENTS FOR SOA

In the next two sub-sections, we outline the principles involved in the design of authorization framework for web services and business processes layers of the SOA.

A. Authorization Principles for Web Services Layer

(i) Support for multiple access control models — Authorization service must be able to support a range of access control models. This is necessary because it is not realistic to expect every web service-based application to use the same access control model. In fact, where web services

are used to expose the functionality of legacy enterprise applications, it is likely that organizations will prefer to use their currently existing access control models and mechanisms that they have been using, before exposing the legacy applications as web services. Therefore, authorization architecture must be flexible enough to support multiple access control models including the traditional Discretionary Access Control (DAC), Mandatory Access Control (MAC), Role Based Access Control (RBAC) and the Capability/Certificate-based Access Control models [7, 8].

(ii) Authorization Policies — Languages have long been recognized in computing as the ideal vehicles for dealing with the expression and the structuring of complex and dynamic relationships. Over the recent years, a language-based approach to specifying access control policies has (rightly) gained prominence; this is helpful for not only supporting a range of access control policies but also in separating the policy representation from policy enforcement. Hence an important design principle is to enable the support for a range of policy languages for specifying authorization policies. The policy language(s) used may support fine-grained and/or coarse grained authorization policies depending on the organization's requirements.

(iii) Authorization Credentials -- It is necessary to define what access-control related credentials are required and how to collect them. Some access control mechanisms may pull the credentials from the respective authorities and send them to the responsible authorization components. For example, in the semantic approach [9], the AC Proxy component collects the relevant privilege (attribute) certificates (for the client) from the PMI Client component which in turn requests the appropriate PMI Node for the privilege certificates for the client. Other access control mechanisms may expect the client to collect the credentials from the respective authorities and push them to the responsible authorization components. For example, [2] proposes a model in which a client itself collects the required authorization credentials from the relevant authorities and sends the set of credentials collected before invoking a web service. Hence, we recommend an authorization architecture designed for web services to be able to support both the push and pull models of collecting credentials.

(iv) Decentralized and Distributed Architecture — Given the distributed decentralized nature of the web, it is reasonable to demand that an authorization architecture designed for web services should embrace the same decentralized nature. As an example, an organization may typically have a hierarchical internal structure. The decentralized approach allows us to specify authorization policies for web services on an organizational unit level for different components in the web service hierarchy. A distributed architecture provides many advantages such as fault tolerance and better scalability, and can outweighs its disadvantages such as more complexity and communication overhead.

The WSAA architecture described in [6] has taken these principles into account in its design and management.

B. Authorization Principles for Business Process Layer

(i) *Decentralized Policy Administration* — Each partner involved in a business process workflow should be allowed to control its set of authorization policies autonomously whether the partners are from within one organization or from multiple organizations. The authorization architecture must not place a constraint on coordinating the policies across different domains involved in the virtual enterprise [4] formed by a business process workflow.

(ii) *Dynamic Discovery of Business Process Partners' Authorization Evaluation Components* — Every web service must let its potential clients be aware of the access control mechanism it uses and where, if necessary, to get the credentials from and where to send them for making the authorization decision. This could be achieved in the form of the assertions in the WS-AuthorizationPolicy (defined in [6]), similar to the WS-SecurityPolicy specification) statements. In the case of a dynamic business process, where the binding to actual implementations of partner web services is made at runtime using some pre-established criteria, the authorization architecture must make the client aware of each of the partners' authorization mechanisms and components involved. A coordination component may be used to send such information to clients. When the flow reaches a stage where some credentials are required by the access control system of the partner involved, the coordination component can make the client aware of what authorization credentials to send to the partner's component for the authorization evaluation to be made.

(iii) *Orchestration of Partners' Authorization Evaluation Components and Combination of Individual Decisions* — Authorization architecture must use some form of coordination mechanism to orchestrate the partners' authorization evaluation components and the client involved in a workflow. In the case of dynamic¹ business processes, a coordinator, for instance, should maintain session state information so that all or some partners know the authorization given to a client. As in any complex transactions handling mechanism, there must be a mechanism in place to either commit or rollback an authorization decision based on the authorization decisions from partners' policy evaluators. A business process may be performed only when all the partners' authorization components involved give out a positive authorization. Decision-combination algorithms such as those defined in the RAD architecture [10] should be defined by the partner controlling the workflow to combine and give out a final authorization decision.

¹ In a dynamic business process, only the partner interfaces are defined at the design time, but not the actual bindings to real instances of partner services.

(iv) *Non-disclosure of Policies* — A partner or a set of partners involved in a business process may not wish to disclose their policies to the partner that is controlling the business process. It is an important requirement that the authorization architecture should not need all the partners to disclose their policies to other partners involved in the workflow. For example, if a Travel Agent Service (TAS) creates a business process that binds and interacts with United Airlines, Hertz car rental and Hilton Hotel at design or runtime, the TAS should not require the different partners involved to disclose their policies to manage the authorization decisions involved. Large organizations would want to set and enforce their policies themselves or by outsourcing to a trusted partner (who runs their authorization service). However, they would want to do business by binding to portal travel agents using a secure SOA.

III. BUSINESS PROCESS AUTHORIZATION ARCHITECTURE

Before we delve into the design of the architecture, we clearly distinguish between *static* and *dynamic* business processes. A static business process is a pre-composed business process, where all the partner service interfaces and their binding information are known at design time itself. A dynamic business process is more complex, where only the partner interfaces are defined at design time, but not the actual bindings to real instances of partner services (web services and/or business processes). The binding is made at runtime to real instances of services by letting the client interact with the business process. For instance, a travel agent may statically bind at design time to always book (i) flight tickets with United Airlines, (ii) cars with Hertz car rental and (iii) hotel rooms at the Hilton. But in real-world situations, customers want more flexibility; and therefore, travel agents may opt to expose their services as dynamic business processes, where the customer at runtime chooses an appropriate partner service (such as airline, car rental agency, or hotel) depending on their own requirements. We make an important assumption in this paper. A dynamic business process may not only invoke partner web services but also partner services that are themselves business processes.

A. BPAA Architecture Design

The proposed architecture is shown in Figure 3. The BPAA comprises an administrative domain and a runtime domain. We manage business processes in the administration domain. Authorization related components such as authorization policy evaluators, certificate and credential authorities and dynamic attribute services can be managed in the administration domain. Also security administrators can assign a set of authorization policy evaluators to authorize requests to business processes. We have a runtime domain where the authorization related information such as what credentials are required to invoke a particular business process and how to collect those credentials is compiled and stored. This makes the authorization process efficient. This information is automatically compiled from time to time when necessary by using the information from the

administration domain and it can be readily used by components in the runtime domain. A client makes use of a registry server such as a UDDI directory to find business process definitions (WS-BPEL statements).

Let us now consider the various system components involved in the BPAA. The components *Authorization Policy Evaluators*, *Certificate and Credential Authorities*, *Dynamic Attribute Services*, and *Authorization Decision Composers* are system objects in our architecture. The *Authorization Manager (AZM)* for an organization is responsible for managing these components. The *Authorization Administration API* is used to manage these components and the related data is stored in the *Authorization Administration Database (AAD)*.

The *Certificate and Credential Authority (CCA)* is responsible for providing authentication certificates and/or authorization credentials required to authenticate and/or authorize a client. For example, a CCA may provide authentication certificates such as X.509 or authorization credentials such as a Role Membership Certificate (RMC) or a Privilege Attribute Certificate (PAC). We define Certificate and Credential Authority as a tuple, $cca = [i, l, CR, pa, ra(pa)]$, where i is a URN, l is a string over an alphabet Σ^* representing a network location such as a URL, CR is the set of authentication certificates and/or authorization credentials cca provides, pa is an input parameter representing a subject, ra uses pa and gives out an output (result) that is the set of certificates/credentials for the subject.

The *Dynamic Attribute Service (DAS)* provides system and/or network attributes such as bandwidth usage and time of the day. A dynamic attribute may also express properties of a subject that are not administered by security administrators. For example, nurses may only access a patient's record if they are located within the hospital's boundary. A DAS may provide the nurse's 'location status' attribute at the time of access control. Dynamic attributes' values change more frequently than traditional *static* authorization credentials (also called privilege attributes). Unlike authorization credentials, dynamic attributes must be obtained at the time an access decision is required and their values may change within a session.

We define Dynamic Attribute Service as a tuple, $das = [i, l, AT, pd, rd(pd)]$, where i is a URN, l is a string over an alphabet Σ^* representing a network location such as a URL, AT is the set of attributes that das provides, pd is input parameter(s) representing attribute(s) name(s), rd uses pd and gives out an output (result) that is the value of the attribute(s).

The *Authorization Policy Evaluator (APE)* is responsible for making authorization decision on one or more abstract system operations. An APE may use a type of access control mechanism and an authorization policy language that may be unique to it. However, we define a standard interface for the set of input parameters an APE expects (such as subject identification, object information and the authorization credentials) and the output authorization result it provides.

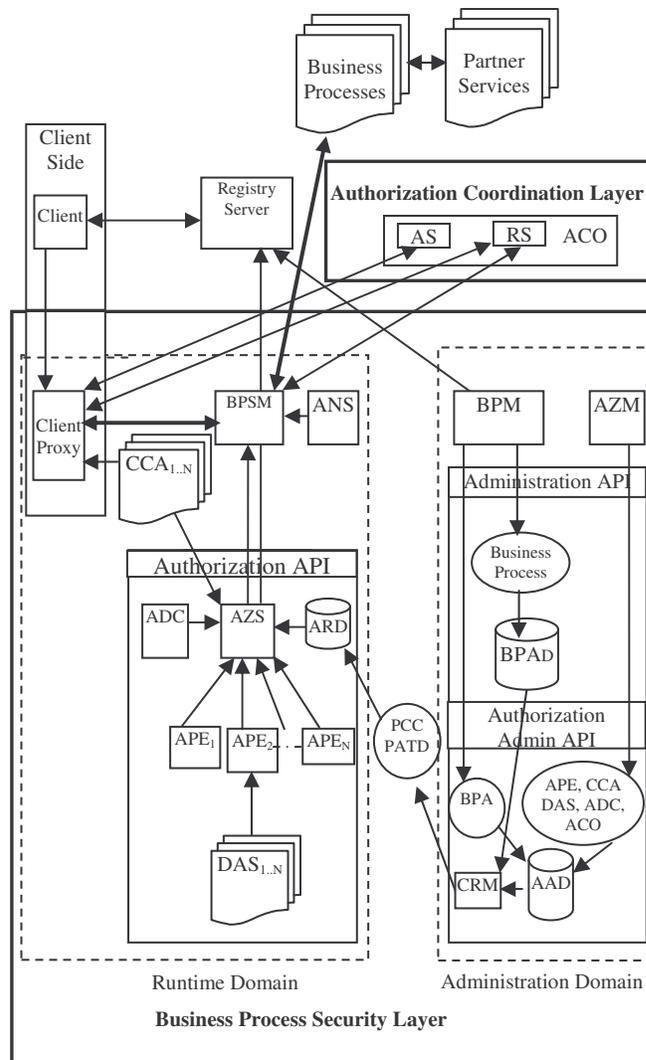


Fig. 3: BPAA Overview Diagram

We define Authorization Policy Evaluator as a tuple, $ape = [i, l, pe, re(pe), OP, DAS, CCA]$, where i is a URN, l is a string over an alphabet Σ^* representing a network location such as a URL, pe is the set of input parameters such as subject and object details, re is a function that uses pe and gives out an output (result) of authorization decision. OP is the set of abstract system operations for which ape is responsible. The DAS is the set of dynamic attribute services responsible for providing dynamic runtime attributes to the ape . The ape uses these attributes to make authorization decisions. The CCA is the set of certificate and credential authorities that provide the credentials required by the ape .

The *Authorization Decision Composer (ADC)* combines the authorization decisions from various authorization policy evaluators involved by using an algorithm that resolves the authorization decision conflicts and combines them into a final decision. We define Authorization Decision Composer as a tuple, $adc = [i, l, a, pc, rc(pc)]$, where i is a URN, l is a string over an alphabet Σ^* representing a network location such as a URL, a is the name of a pre-defined algorithm adc

uses to combine the decisions from the individual authorization policy evaluators. The pc is an input parameter representing the decisions from individual authorization policy evaluators, rc uses pc and the authorization decision composer algorithm a to combine the decisions and gives out an output (result) that is the value of the final authorization decision.

The runtime domain consists of the Client Proxy, the Business Process Security Manager, the Authentication Server, the Authorization Server, and the Authorization Coordinator components.

The *Client Proxy (CP)* collects the required authentication certificates and/or the authorization credentials from the respective authorities on behalf of the client before sending a request to a business process and handles the session on behalf of the client with a *Business Process Security Manager*.

The *Business Process Security Manager (BPSM)* is responsible for both the authentication and the authorization of the client to a business process. A client's Client Proxy sends the necessary authentication certificates and authorization credentials to the BPSM. It is responsible for managing all the interactions with a client's Client Proxy.

The *Authentication Server (ANS)* receives the authentication certificates from the BPSM and uses a mechanism to authenticate the client. We treat the ANS as a black box in our architecture as our focus in this paper is on the authorization of the client. We included this component in the business process security layer for completeness.

The *Authorization Server (AZS)* decouples the authorization logic from the application logic. It is responsible for locating the business process' Authorization Policy Evaluators, sending the credentials to them and receiving the authorization decisions. Once all the decisions come back, it uses the business process' Authorization Decision Composer to combine the authorization decisions. If required, the AZS also collects the required authorization credentials on behalf of clients from the respective Certificate and Credential Authorities.

The *Authorization Coordinator (ACO)* is used to coordinate the authorization between a client (by involving the Client Proxy) and the *dynamic* business processes and their partner services (web services and/or business processes). It is composed of an *Activation Service (AS)* and a *Registration Service (RS)* that expose standard interfaces to the participants (Client Proxy and Business Process Security Manager) in the authorization coordination protocol.

The *Business Process Managers (BPMs)* manage a set of business processes for which they are responsible in an organization. They use the Administration API shown in Figure 3 to manage the business processes. The business process definitions are stored in the *Business Process Administration Database (BPAD)*, see Figure 3.

We define a Business Process as a tuple, $bp = [i, l, \Sigma, WS, BP, B, pa, MD, bpm, bpsm, aco]$, where i is a non-empty string over an alphabet Σ^* representing a globally unique identifier such as a URN, l is a string over an alphabet Σ^* representing a network location such as a URL, Σ is a finite set of states representing the internal state of the business

process at a given time, WS is the set of URNs of partner web services or activities that comprise the business process, BP is the set of URNs of partner business processes or activities that comprise the business process, B is the network protocol binding such as SOAP over HTTP for the business process, pa represents the business process flow algorithm represented in a WS-BPEL statement, MD is the metadata providing additional description for bp , the bpm is the identity (ID) of the Business Process Manager (BPM) responsible for managing bp . The $bpsm$ is the location of the Business Process Security Manager component responsible for the authentication and authorization of the clients to the business process. The aco is the location of the Authorization Coordinator responsible for coordinating the authorization of a client to the bp 's partner services. The aco is defined only for dynamic business processes and is null for static business processes.

The Σ , B , or the MD can be the empty set \emptyset . If B is an empty set, \emptyset , then the business process defined is either an *abstract* business process or a *dynamic* business process. An abstract business process is not executable and only defines the standard interfaces between a business process and its partner services and the messages passed between them. If it is a dynamic business process, the individual bindings to partners are made at runtime using the client's preferences. If B is not an empty set at business process design time, then it is a static (pre-composed) business process.

B. BPAA Authorization Policy Evaluation

The Business Process Managers (BPMs) are also responsible for managing the authorization-related information for the business processes for which they are responsible. This information is stored in the *Business Process Authorization* tuple, $bpa = [i, bp, APE_{bp}, adc_{bp}]$, where i is a URN, bp is the business process to which bpa is defined. The APE_{bp} is the URNs of the set of Authorization Policy Evaluators responsible for authorizing the requests from a client to the bp . The adc_{bp} is the URN of an *Authorization Decision Composer*. It is responsible to combine at runtime, the authorization decisions given out by the set of APEs in the APE_{bp} .

At the time of evaluation in runtime domain, Credential Manager (CRM) component in the BPAA is responsible for compiling and storing the authorization information required by the components. This runtime authorization information is stored in the *Authorization Runtime Database (ARD)* (Fig. 3). The runtime authorization information consists of two tuples namely, BusinessProcess-Credential-CCA tuple (PCC tuple) and BusinessProcess-Attribute-DAS tuple (PATD tuple). The CRM is invoked from time to time, when a business process object is created or modified in the BPAD.

The BusinessProcess-Credential-CCA tuple is defined as $pcc = [i, bp, CR, cca, ape]$, where i is a URN, bp is the URN of the business process, CR is the set of authorization credentials to be obtained from the Certificate and Credential Authority, cca to get authorized to invoke bp . The ape is the URN of the Authorization Policy Evaluator that requires these credentials. This means each bp can have one or more of these (tuple) entries in the ARD.

The BusinessProcess-Attribute-DAS tuple (PATD tuple) is defined as $patd = [i, bp, AT, das, ape]$, where i is a URN, bp is the URN of the business process, AT is the set of attributes to be obtained from a Dynamic Attribute Service, das to make an authorization decision. The ape is the URN of the Authorization Policy Evaluator that requires these attributes. This means each bp can have one or more of these (tuple) entries in the ARD.

C. BPAA Authorization Algorithms

The BPAA supports three authorization algorithms. The first, the *push-model* algorithm, supports the authorizations where a client's Client Proxy (CP), using the information in the BP-AuthorizationPolicy, collects and sends the required credentials (from the CCAs) and attributes (from the DASs) to a Business Process Security Manager (BPSM). The second, the *pull-model* algorithm, supports the authorizations where the Authorization Server (AZS) itself collects the required credentials from the CCAs, and the APEs collect the required attributes from the DASs. The AZS in this case uses the runtime objects information from the Authorization Runtime Database to be able to do so. The third, the *combination-model* supports both the push and pull models for collecting the required credentials and attributes. An organization must deploy one of these algorithms depending on the access control mechanisms used by the business process.

When the combination-model algorithm is deployed by an organization, the organization's Authorization Manager (AZM) may arbitrarily decide whether the credentials required from a CCA and dynamic attributes required from a DAS for each of the business process' APEs are fetched by a Client Proxy (push-model) or by the authorization components themselves (pull-model). The AZM may decide to give the entire responsibility of fetching the required credentials and attributes to the client proxy or to authorization components or share responsibility of fetching credentials and attributes amongst the client proxy and the authorization components. This information is reflected in a business process' BP-AuthorizationPolicy.

The BP-AuthorizationPolicy includes assertions that specify what credentials (and from which CCA) and attributes (and from which DAS) a client's Client Proxy has to collect before invoking a business process. These assertions also include the credentials and attributes required to invoke a *static* business process' partner Web services as well as its partner business processes. We extend the WS-BPEL statement schema to include the BP-AuthorizationPolicy. Note that the partner Web services and business processes-related authorization information is not included in the BP-AuthorizationPolicy of a dynamic business process. Such information is only necessary for a static business process. Finally, the authorization coordination information is also included in the BP-AuthorizationPolicy. This information is necessary only for the dynamic business processes. We have designed and implemented the authorization evaluation schemes for both static and dynamic business processes in push, pull and combined model scenarios. Due to lack of space in this

paper, we refer the reader to [11] where they are described in full.

D. Dynamic Business Process Authorization

We leverage the WS-Coordination framework [12] to coordinate authorization of a client to a dynamic business process and its partner services. When a client invokes a dynamic business process, the Client Proxy component is responsible for the activation of a new instance of an Authorization Coordinator. It is aware that authorization coordination is required to get the client authorized to a dynamic business process, because the BP-AuthorizationPolicy has the information about the authorization coordinator, the coordination protocol used and its type (authorization coordination type), and finally its location. The Business Process Security Manager is another participant in the coordination protocol. During the course of execution of a dynamic business process, if the WS-BPEL Engine needs to invoke a partner service, it sends a message about the same, to the business process' BPSM. BPSM then informs the Authorization Coordinator that a partner service has been invoked and it needs authorization credentials from the client (Client Proxy). The Authorization Coordinator also informs the Client Proxy about the same. The Client Proxy fetches the required credentials and gets back to the Authorization Coordinator. The Authorization Coordinator then sends a message with the received credentials to the Business Process Security Manager. BPSM sends these credentials to the BPEL Engine. The BPEL engine uses these credentials and then continues execution of the partner service.

The Client Proxy interacts with the BPSM sending and receiving messages as normal, with the exception that it embeds the authorization coordination context (which carries the authorization information) in a SOAP header block in its messages to provide authorization credentials for those partner services (web services and/or business processes) that are invoked. Also the Client Proxy itself registers as a participant with the authorization coordinator. The BPSM understands the protocol messages associated with our authorization service. If it has not registered a participant previously, it does so once it receives a SOAP message from the Client Proxy containing an authorization context header using the details provided in the context (via the WS-Coordination registration service URI). This register operation occurs every time that the BPSM receives a particular context for the first time.

When the Client Proxy receives the final response from the BPSM after the execution of the business process, it sends a *Completion Message* to the Authorization Coordinator. The Authorization Coordinator then sends the Completion Message to the BPSM registered as a participant to the Authorization Coordinator. Any subsequent calls by the Client Proxy (on behalf of the client) to that business process with the same context will result in the service being unable to register a participant since the context details will no longer resolve to a live coordinator with which to register. This is shown in Figure 4.

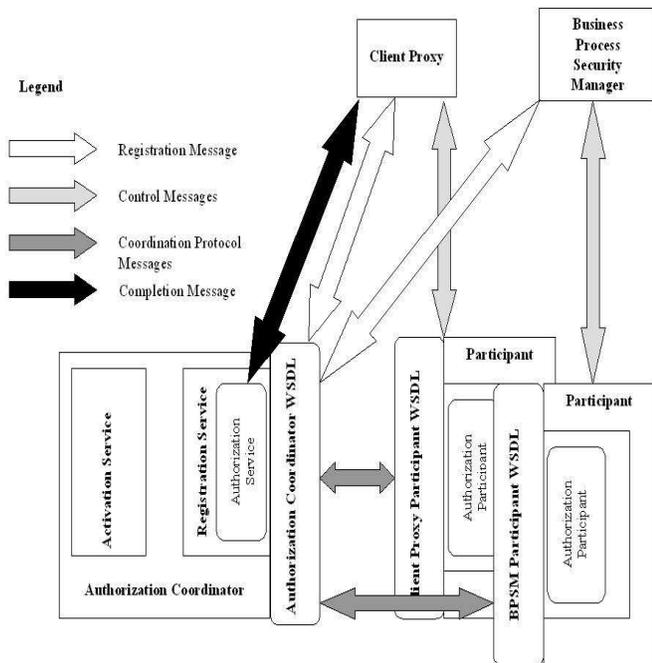


Figure 4: Authorization Coordination Framework

We have developed an implementation of this architecture using Microsoft BizTalk Server to create sample business processes in the BPEL4WS and demonstrated the features of the BPAA using those business processes. The architecture was used to demonstrate a healthcare application scenario using .NET framework. Performance evaluation showed that the BPAA architecture introduced an average performance delay of some 200ms when invoked with different processes.

IV. CONCLUDING REMARKS

In this paper, we have proposed and developed authorization architecture for business processes (BPAA) in SOA. Our authorization architecture is able to support both static and dynamic business processes. Also, a business process may have web services or even other business processes as partners. We took all such scenarios into consideration and have provided a comprehensive architecture for authorization for the business process layer of the SOA. Also we extended our authorization coordination framework to allow for both static and dynamic business processes to invoke partner services that are themselves dynamic business processes. The proposed BPAA supports multiple access control models including the MAC, DAC, and the RBAC models. Access control mechanisms can either use push or the pull model or even a combination of both, for collecting client credentials. Our architecture provides decentralized security administration. The partners involved in a business process workflow are allowed to autonomously control their authorization policies. The partners can be either from within an organization or from multiple organizations. In the case of static business

processes, the information about the authorization credentials required to invoke the partner services is exposed in WS-BPEL using BP-AuthorizationPolicy at the design time itself. In the case of dynamic business processes, dynamic discovery and orchestration of business process' partners' authorization evaluation components are achieved. The BPAA coordinates the authorization where binding to real partner services happens at runtime depending on client requirements. Furthermore, The BPAA does not require the partner services to disclose their policies to the partner that is controlling the business process. The authorization of the client happens at the same place, where the partners originally intended it to be. Hence organizations can now leverage the services offered by the BPAA and do business by binding to the portal agents even if they do not trust them to perform client authorization. The Business Process Security Manager can be placed in a firewall zone, which enhances the security of business processes placed behind an organization's firewall. We have implemented the proposed architecture and its components using the .NET middleware platform and demonstrated its operation by developing a healthcare application scenario over this architecture.

REFERENCES

- [1] T. Andrews et al., Business Process Execution Language for Web Services, <http://www.ibm.com/developerworks/library/specification/ws-bpel/> (accessed Jan 2011).
- [2] S. Agarwal, B. Sprick, and S. Wortmann, "Credential Based Access Control for Semantic Web Services," American Association for Artificial Intelligence, pp. 110-120, 2004
- [3] R. Kraft, "Designing a Distributed Access Control Processor for Network Services on the Web", Proc of the ACM Workshop on XML Security, USA, pp. 36-52, 2002.
- [4] H. Koshutanski and F. Massacci, "An Access Control System for Business Processes for Web Services," Informatica e Telecomunicazioni, University of Trento, Technical Report DIT-02-102, 2002
- [5] M.C. Mont, A.Baldwin and J.Pato, "Secure Hardware-based Distributed Authorization underpinning a Web Service Framework", HPLabs Technical Report HPL-2003-144, 2004.
- [6] S. Indrakanti, V.Varadharajan and R.Agarwal, "On the Design, Implementation and Application of an Authorization Architecture for Web Services", International Journal for Information and Communication Security, Vol.1, No.1/2, pp. 64-108, 2007.
- [7] D.W. Chadwick and A.Otenko, "The PERMIS X.509 role based privilege management infrastructure", Future Gener. Comput. Syst. 19, pp. 277 - 289, 2003
- [8] V. Varadharajan, C.Crall and J.Pato, "Authorization in Enterprise wide Distributed Systems: Design and Application", Proceedings of the 14th IEEE Computer Security Applications Conference, pp. 178-189, 1998.
- [9] M.I. Yague and J.M.Troya, "A Semantic Approach for Access Control in Web Services". In Euroweb 2002 Conference. The Web and the GRID: from e-science to e-business, Oxford, UK, pp. 483-494, 2002.
- [10] K. Beznosov et al., "A Resource Access Decision Service for ORBA-Based Distributed Systems," in Proceedings of the 15th Annual Computer Security Applications Conference: IEEE, pp.310-319, 1999.
- [11] S. Indrakanti, "Engineering Authorization Services for Service Oriented Architectures, PhD Thesis, Macquarie University, 2007.
- [12] L.F. Cabrera et al., "Web Services Coordination Framework", <http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-tx/WS-Coordination.pdf> (accessed Jan 2011).

Trust Metrics for Services and Service Providers

Zainab M. Aljazzaf, Mark Perry
 Department of Computer Science
 University of Western Ontario
 London, Canada
 {zaljazzaf, mperry}@uwo.ca

Miriam A. M. Capretz
 Department of Electrical and
 Computer Engineering
 University of Western Ontario
 London, Canada
 mcapretz@uwo.ca

Abstract—Trust is as significant a factor for successful online interactions as it is in offline communities. Trust is an important factor that is used as a criterion for service selection. There is a need to know information about services and service providers to establish trust and identify their trustworthiness. Most trust studies focus on trust establishment for services without clearly identifying trust information for services and service providers. Services and service providers traverse many domains with different properties and requirements. Identifying a unified trust information (trust metrics) for such an open environment is a challenge. This paper proposes a unified trust metrics classification for services and service providers. The proposed trust metrics can be extended and used in an open environment or within specific domains to establish trust for services and service providers.

Keywords - Trust; trust metrics; service; service provider.

I. INTRODUCTION

In human communities, there is uncertainty about the behaviour of strangers. People avoid interacting with others who they do not trust. Trust plays a significant role in facilitating the interaction in such uncertain environments. A *Truster* is the subject that trusts a target trusted entity known as a *Trustee*. We define trust as *the willingness of the trustor to rely on a trustee to do what is promised in a given context, irrespective of the ability to monitor or control the trustee, even though negative consequences may occur* [1].

Building a distributed software system requires the interaction and use of resources from diverse organisations throughout the Web. In such diverse systems, different entities spread around different domains and organizations, and pass the boundary of a particular physical community, which may have clear security and trust preferences. Service Oriented Architecture (SOA) is “an architectural style for building enterprise solutions based on services” [2]. There are three roles in SOA as shown in Figure 1 [3]: *service provider*, an organization or platform that owns, implements, and controls access to the services; *service requestor*, an application, services, or the client who is looking for and invoking a service; and *service registry*, a searchable directory where the description of the services is published by the providers and searched by the requestors.

There are many services with similar functionalities. The non-functional properties of a service can be a differentiating factor between the similar services and as a criteria for service selection. Quality of Service (QoS) is the quality aspect of a

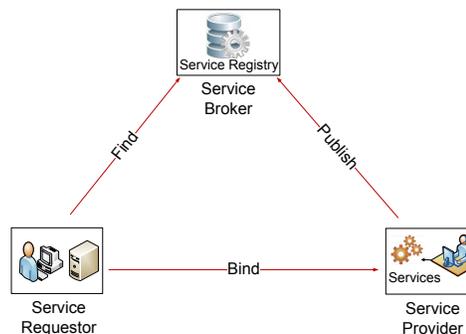


Fig. 1. Services roles and operations in SOA [3].

service [4], and is considered as a non-functional property of a service. Trust has been used as a criteria for service selection [5][6][7][8]. The trustworthiness of a service is considered as a non-functional property of a service. Service requestor (trustor) may select a service/provider (trustee) based on its trustworthiness.

The trustworthiness of a service provider can enhance the requestor’s trust in its services [9]. A requestor can select a service from providers of the highest level of trust [10]. Considering trustworthiness of service providers supports trust bootstrapping (rating new comers) the providers’ new services. For example, if a provider is known to be a trustworthy, requestors will trust the provider’s services and encourage to select its new services. Therefore, it is important to establish trust for service providers and select a service based on its provider’s trustworthiness in addition to the service’s own trustworthiness.

Trust is based on information [1], but it is difficult to determine the information that should be used. In the offline world, traditional forms of communication allow people to assess a wider range of cues related to trustworthiness than is currently possible through online communication. The Internet gives little evidence about the solidity of the entity behind it. The challenge is to find sufficient online substitutes for the traditional cues to trust, which are obvious in the physical world and identify new information elements, which are appropriate for deriving measures of trust [11].

Trust Metrics (TM) is a new term and is defined in this paper as *the information of an entity that is required and used to evaluate the trustworthiness of the entity*. An entity, in this work, can be a service or service provider. TM is the first party (i.e., a service or provider) information provided by a service or service provider to evaluate its trustworthiness [1]. For example, a service can present its reliability as a trust metric for the requestors to trust the service based on the reliability trust metric. To identify trust metrics, it is important to explore what information is required to build trust for services and service providers.

Information has many dimensions and each service/provider sets its own information. In SOA, transaction may span a range of domains and organization. Services and service providers may traverse many domains with different properties and requirements. For example, a requestor of a service has many requirements and each seeks for different services' properties. Therefore, a domain may need to support a range of trust metrics and this requires to identify a unified trust metrics for such an open environment. Some studies try to overcome this problem by defining a notion of community [12] or address trust in specific domains [5][13][14]. This paper proposes a unified trust metrics classification for services and service providers that is suitable for SOA environment.

The rest of the paper is organised as follows: Section II presents related work. Trust metrics and trust principles is presented in Section III. Section IV presents trust metrics and QoS. The proposed trust metrics for services and service providers is presented in Section V. Section VI concludes the paper.

II. RELATED WORK

In the literature, there is no clear identification of trust metrics for services. In addition, there is no defined trust metrics for service providers.

A. Trust Services

Zhengping et al. [15] defined domain specific trust information that is limited. The work monitors the behaviour of a trusted Web Service in case it has bugs during operation, which will drop the trust degree of the Web Service. The authors define properties to establish trust for services such as functions and run time environment, and for recommender who recommend a service such as popularity and authenticity of the description. Different domain characteristics defined by the system analyst. Kim and Doh [16] propose the selection of the optimal path to compose a number of Web Services based on QoS information and trust type (the computed trust level based on aggregated ratings from the consumer of the services, which indicate the estimation of the reliability of the service provider). The authors assume that trust type is associated with each service where the assignment of trust types performed by the clients themselves or trust authority. Trust metrics are not specified and trust is based on assumed trust type.

Maximilian and Singh [10] made a distinction between trust and QoS and presented the selection of a Web Service based

on non-functional attributes such as QoS and trust. Wang and Vassileva [9] stated the importance to define information needed for trust and reputation mechanism. They stated the use of QoS to build trust where trust and reputation are built for each quality property of a service and the overall trust and reputation depend on the combination of trust and reputation for each property.

Other researchers address trust as a QoS [5][6][8][17], build trust based on a set of QoS [8][17][18][19], or build trust based on a set of QoS related to specific system, application, or domain [5][17][18][19]. Dragoni [5] mentioned that evaluation of trust is a key QoS aspect of Web Service selection. The author used security features of the service to establish trust (satisfying the provider's trust security requirements). Ying-Feng and Pei-Ji [6] specify trust or reputation as one of the QoS of Web Service. Kalepu et al. [8] identified a new QoS attribute, *verity*, as an important contributor to the quality driven selection and composition of Web Services and to be a measure of trustworthiness of a Web Service. Verity refers to the degree of variance in the compliance levels of the services and assesses the reputation of the provider based on local and global rating. They identify verity for Web Services and verity for Web Services providers. However, trust is not a QoS and there is a clear distinction between the two terms.

In [18], reputation is modelled as a vector of QoS attributes such as performance and reliability. Jin-dian et al. [19] establish trust based on whether it is secure enough to access a service or how to choose a more reliable provider. They measure the possibilities of providing cheating or malicious behaviour and satisfaction values to measure how satisfied a user feels about a given interaction (both are real numbers in [0,1] and a high rate reflects a high interaction quality). The trust evaluation can take many aspects (QoS requirements) into account such as process time and access speed. Vu et al. [17] rank services according to its prospective level of satisfying user's QoS requirements. However, building trust should consider other properties beside QoS.

B. Trust Service Provider

In Web Services and SOA, the idea of trusting a service based on its provider is neglected [9]. Trust in the Internet has a clear distinction between the two and has identified quality requirements for providers to assist their trustworthiness and help users in their decision to use providers' services [20][21].

Jin-Dian et al. [19] presented the idea of assigning trust provider rate to its new Web Services. They mentioned that assigning trust rate to the provider is an interesting research problem. They stated that a registry that has past experiences with the Web Service's provider initializes the rate of the new Web Service to be equal to its provider's rate.

The work in [18] assesses the trustworthiness of a Web Service provider by measuring its reputation based on the rate given by the user. However, identifying trust information supports the trust bootstrapping process (i.e., rating new services and service providers). Maximilien and Singh [10] mentioned that if service provider is already determined to be

trustworthy, then the selection of services will be based on their provider's rating. The authors stated that determining the trust level to be assigned to providers is a nontrivial process. However, identifying trust metrics helps to establish trust.

III. TRUST METRICS AND TRUST PRINCIPLES

In our previous work [1], we identified the trust principles. This work follows the trust principles to identify TM. The TM addresses the following trust principles :

- Trust and risk: Requestors have no control over services that advertise only their interface. Less perception of control increases the risk. Under a risky exchange situation it is important to include penalties, rewards, insurance, and other risk remedies in case something goes wrong. Risk remedies can be identified as a TM.
- Trust development phases: Trust goes through three development phases: trust building, stabilising trust, and dissolution [21]. Most studies assume a system where trust and reputations already exist (i.e., stabilising trust phase). In trust building phase, it is important to initialize a trust rate for a new service or a new service provider. Identifying TM is important in the initialization process, where the TM are rated and the overall trust for a service and service provider can be the average of trust for each TM.
- Trust relationship properties: Trust is usually specified in terms of a relationship between a trustor and a trustee. Trust relationship can be one-to-one between a requestor and a service and one-to-many between a requestor and a group of services (i.e., a provider who provides a group of services). By identifying trust metrics the system can support the context specific characteristic of trust (trust a service to perform a specific action within a specific context), where a requestor can select a service based on a set of trust metrics.
- Trust is based on information. There is a need to know information about the services and service providers to establish trust.
- First party information: First parties (i.e., services/service providers) should provide the information to develop trust. For example, QoS properties and other information (e.g., delivery methods, insurance, privacy, security, pricing, and availability) can be considered as important information on which to build trust.
- The distinction between trust and QoS: Trust is not a QoS aspect of a service or a service provider. There is a clear distinction between the two terms' definitions as presented in the introduction. QoS properties can be identified as TM to establish trust.
- Security and privacy: Security and privacy are important factors to consider in the trust establishment process. Security and privacy can be considered as important TM.
- Provider's trustworthiness: Trust ratings of a service and its provider are related and each one affects the other. Therefore, it is important to identify providers' trustworthiness and define TM for service providers.

IV. TRUST METRICS AND QoS

QoS can be identified as an important TM to establish trust. This work defines QoS as TM. To identify TM for an open system, it is important to generalize a list of TM applicable for most services. As a part of generalization process, it is required to generalize QoS for diverse services and service providers. To define a unified TM classification, we need to extract diverse QoS from the literature. Some QoS can be measured and some are not. It is important to include and quantify the non-measurable QoS to be used in trust rating algorithm and calculation.

There are many research efforts to define and categorize QoS and how to express, quantify, and model them [4][16][22][23]. In [4][8][9][22][24][25][26][27][28][29] generic and business QoS requirements for services are presented. Lee and Shin [26] define a set of major Web Services' QoS attributes. Menasce [28] presents the QoS issues in web services. Yu et al. [23] present a list of QoS and how to calculate each. They specify that security is not quantifiable QoS but they present a formula to test the security of Web Services based on the number of attacks detections. Rahman and Meziane [27] present five essential QoS requirements based on the most used QoS from the literature. These are: readiness, transaction, reliability, speedy, and security.

O'Brien et al. [24] define other QoS requirements for SOA such as: modifiability, testability, and usability. Ran [22] identified other QoS, which are: supported standard, stability/change cycle, and completeness. In addition, there are a domain or application specific QoS. Hoyle [29] identifies other quality characteristics for services such as courtesy, comfort, competence, credibility, dependability, efficiency, effectiveness, flexibility, honesty, promptness, responsiveness. Larson [30] identified serviceability and user satisfaction as performance measurement of service delivery.

Moorsel [25] discusses quantitative metrics and a framework for evaluating internet services. Three metrics are defined that should be emerged to evaluate Business to Consumer (B2C), Business to Business (B2B) and service provider systems. The metrics are: QoS, Quality of Experience (QoE), and Quality of Business (QoBiz). QoE and QoBiz are claimed to quantify the user experience and business return, respectively.

Based on the aggregated QoS in the literature, we propose a classification of QoS into objective QoS and subjective QoS, as shown in Figure 2. Objective QoS are the QoS that can be measured. Subjective QoS are the QoS that cannot be measured. This classification helps to define and classify TM.

V. TRUST METRICS

In this work, TM overcome other trust information in the literature to include information of diverse domains (government, online marketing, bank, etc), QoS, and different possible services and service providers' information and properties. Some TM may not be applied to all services and it is possible to add other TM. Figure 3 shows the proposed TM for services and service providers which is classified into Services Trust Metrics (STM) and service Providers Trust Metrics (PTM).

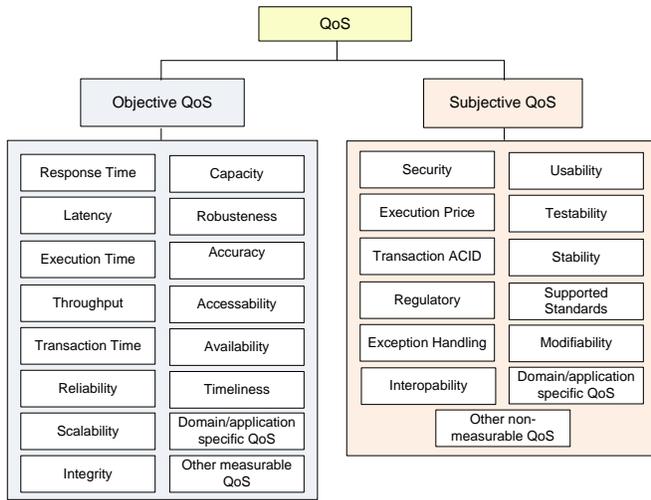


Fig. 2. A classification of QoS.

A. Trust Metrics for services (STM)

STM is classified into Objective STM (OSTM) and Subjective STM (SSTM) as follows:

1) *Objective Services' Trust Metrics (OSTM)*: OSTM are the TM that can be measured. OSTM for services are their objective QoS properties such as response time, latency, execution time, throughput, reliability, domain specific measurable properties, and other services' measurable properties. In the following, a number of OSTM is presented:

- Execution Time OSTM, $OSTM_e(s)$: Is the time taken by a service to execute and process its sequence of activities.
- Latency OSTM, $OSTM_l(s)$: Is the delay time between sending a request and receiving the response, i.e., the time the message needs to reach its destination.
- Response Time OSTM, $OSTM_r(s)$: Is the time required to process and complete a service request.
- Throughput OSTM, $OSTM_{thp}(s)$: Refers to the number of requests a service can process per unit of time.
- Availability OSTM, $OSTM_{Av}(s)$: Is the probability that a service is up and accessible to use.
- Reliability OSTM, $OSTM_R(s)$: Refers to the ability of a service to perform its function correctly with either 'no fail' or 'response failure to the user'. It is related to $OSTM_{Av}(s)$.

2) *Subjective Services' Trust Metrics (SSTM)*: SSTM are the TM that are hard to measure directly. SSTM include functional properties, subjective QoS properties, and other properties of services such as remedies, payment satisfaction, output/item satisfaction, delivery satisfaction, domain specific non-measurable properties, and other non-measurable services properties. The following presents some SSTM. Because SSTM are not measurable TM, it is important to quantify them.

- Remedies SSTM, $SSTM_{rem}$: Is the most important metric that should be provided by a service provider for each of its services. Services should provide remedies in

case any thing goes wrong. Each service has different remedies. For example, if the service is shipment service and there was a delay in shipment, lower the shipment price can be offered as a remedy. Another example is that if a service provides a video and the video was slow referred to the subscribed level of a customer, the service should increase the bandwidth for that customer.

- Security SSTM, $SSTM_{sec}$: A requestor can trust a service or service provider based on security. Security is an important factor to be considered in trust establishment.
- Privacy SSTM, $SSTM_{prv}$: A requestor can trust a service or service provider based on privacy. Privacy is an important factor to be considered in trust establishment.
- Payment Satisfaction SSTM, $SSTM_{pym}$: Refers to the degree of the user satisfaction on the offered service based on the payment, if any. For example, do the service charge the user the same or extra amount, do users pay extra unexpected fees, etc.
- Output/Item satisfaction SSTM, $SSTM_{out}$: Refers to the degree of the user satisfaction on the offered service based on the output/item provided. For example, do they get the same output/item they ordered/expected, are they satisfied with the output/item, the quality of the output/item they received, etc.
- Delivery satisfaction SSTM, $SSTM_{delv}$: Refers to the degree of the user satisfaction on the offered service based on the delivery of the item. For example, do they deliver the item on time, do they return the item in case of dissatisfaction, etc.

3) *Trust Metrics Collection for services*: OSTM and SSTM are rated for each service and can be stored in a registry to be used for rating services and service providers. The following is the collected STM in a matrix format. Each row represents a service and each column represents one of the STM (m:s and n:STM). $STM = OSTM + SSTM$

$$STM = \begin{bmatrix} STM_{11} & STM_{12} & \dots & STM_{1n} \\ STM_{21} & STM_{22} & \dots & STM_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ STM_{m1} & STM_{m2} & \dots & STM_{mn} \end{bmatrix}$$

Trustworthy services support remedies, security and privacy; provide high throughput, fast response time, high availability and reliability; provide lower execution, response and latency times. In addition, trustworthy services get high rates for the execution cost, output, payment, and delivery STM.

B. Trust Metrics for service Providers (PTM)

A good provider rate can enhance the requestor's trust in the provider and its services. If a requestor has an alternative to choose between many services from different providers, he can select a provider with a higher trust rate. Rating providers help to encourage providers to behave well, increases the opportunities of the providers to be selected by consumers, encourage competition between providers, influence the economic growth of the providers positively, increase the usage of the Internet technologies such as e-markets, and evolve commerce online.

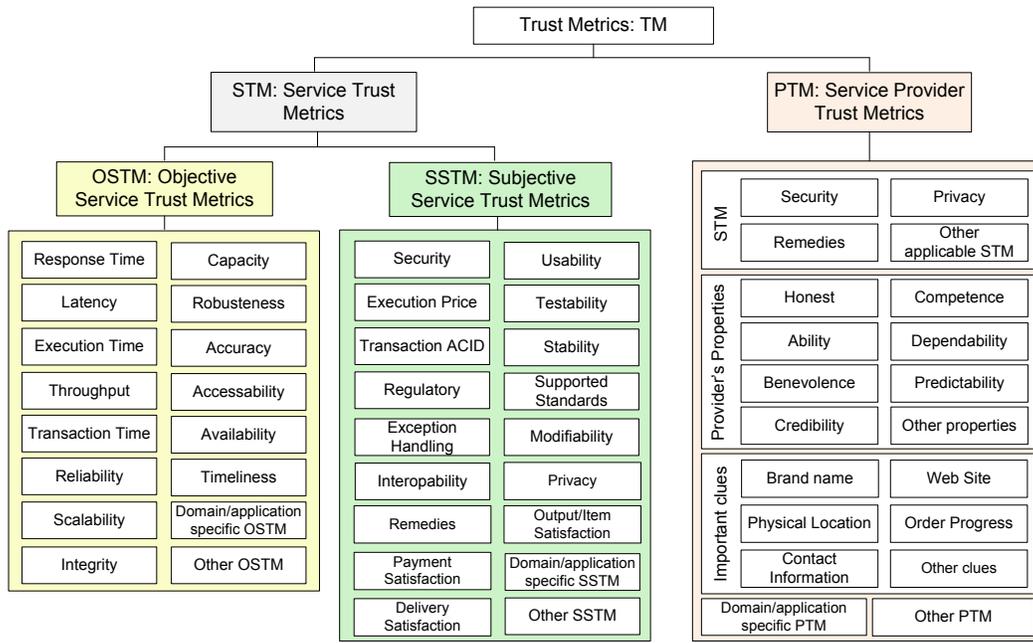


Fig. 3. Trust Metrics.

Trustworthiness of a service provider is based on the *trustworthiness of its services* and the rates of its *properties*. Service providers have many properties that can be considered as useful PTM to build trust. A trustworthy service provider should behave upon its advertised properties and its services advertised properties (i.e., STM). In addition to providers' properties, a provider can provide important clues for requestors to assess its trustworthiness. In the following, a number of important PTM are presented.

1) *Providers' services' properties*: A service provider trustworthiness is based on the trustworthiness of its services. Therefore, STM are implicitly included as metrics to evaluate providers. Any STM can explicitly be identified as PTM to emphasize its importance and can be used as PTM such as security PTM, PTM_{sec} ; privacy PTM, PTM_{prv} ; remedies PTM, PTM_{rem} ; and other applicable STM such as reliability, integrity, robustness, accessibility, availability, timeliness, payment satisfaction, output/item satisfaction, delivery satisfaction, transaction ACID, supported standards, interoperability, and stability.

2) *Provider's properties*: Competence, honesty, ability, benevolence, predictability, credibility, dependability, courtesy, comfort, efficiency, effectiveness, flexibility, promptness, and responsiveness are properties to be considered as PTM. These properties can be evaluated by long term interactions with a provider based on other TM. In the following, competence and honest PTM will be presented.

- Competence PTM, PTM_{comp} : Shows a provider's ability and capability to provide a service and perform the function expected from it (i.e., compliance). Competence is more relevant term for the environment related to

services and computing system [31].

- Honest PTM, PTM_{hons} : The provider that continuously shows its competence will be honest.

3) *Important clues*: Service providers can provide important clues to support their trustworthiness. The more clues a provider provides, the more the provider can support its trustworthiness, and the more is the opportunity for its services to be selected by the requestors. Some clue information may suit some requestors but not others especially if the requestor is an application which dynamically bind to services. The following presents some important providers' clue information, as follows:

- Brand name PTM, PTM_{brand} : A service provider who has a brand name, popular name that is established by a long term interactions with consumers, may encourage the requestors to use its services. A brand name can help in the assessment of service providers' trustworthiness, and this will influence the economic growth of the service providers positively. Trust-based systems can play an important role on the establishment of brand names for service providers. A service provider can provide a name and the system can brand the name based on the level of the trustworthiness of the service provider.
- Web site PTM, PTM_{wsite} : A service provider who has a web site may give an important clue for the requestor to trust the provider and use their services. Web sites may contain information that can assess the trustworthiness of a service provider.
- Contact information PTM, PTM_{inf} : Contact information such as telephone number and e-mail has a great impact in the assessment of the trustworthiness of a service

provider. Having contact information allow the requestors to contact providers to, for example, resolve any issues.

- Retail location PTM, PTM_{loc} : Having physical location, such as physical store, may increase a provider's trustworthiness.
- Order progress PTM, PTM_{ord} : While order progress is more clear offline, it should be provided online, and this may increase a provider's trustworthiness.

4) *Trust Metrics Collection for service Providers*: The evaluated PTM rates for each service provider are stored in a registry to be used for rating purposes. The following is the collected PTM in a matrix format. Each row represents a service provider and each column represents one of the PTM (m:provider and n:PTM).

$$PTM = \begin{bmatrix} PTM_{11} & PTM_{12} & \dots & PTM_{1n} \\ PTM_{21} & PTM_{22} & \dots & PTM_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ PTM_{m1} & PTM_{m2} & \dots & PTM_{mn} \end{bmatrix}$$

Trustworthy service providers support remedies, security, and privacy and provide trustworthy services. In addition, trustworthy service providers provide important clues to support its trustworthiness. By the time, a service provider will become competent and then honest by acting as a trustworthy provider.

VI. CONCLUSION

This paper presents a unified trust metrics classification for services and service providers. Trust metrics cover different information and properties of services and service providers. The trust metrics are extendible and can support domain specific properties. Trust metrics support the trust principles. Each trust metric may require different techniques to gather and evaluate its trust rates. As a next step, trust models to rate the trust metrics, services, and service providers will be established. In addition, there is a need to build a trust framework that establish trust for services and service providers and supports trust-based service selection in SOA.

REFERENCES

[1] Z. M. Aljazzaf, M. Perry, and M. A. Capretz, "Trust online: Definition and principles," *ICCGI 2010: The Fifth International Multi-Conference on Computing in the Global Information Technology*, 2010.

[2] M. Rosen, B. Lublinsky, K. T. Smith, and M. J. Balcer, *Applied SOA: Service-Oriented Architecture and Design Strategies*. Wiley Publishing, 2008.

[3] M. Papazoglou, *Web Services: Principles and Technology*. Prentice Hall, 2008.

[4] K. Lee, J. Jeon, W. Lee, S. Jeong, and S. Park, "QoS for web services: Requirements and possible approaches," W3C, Web Services Architecture Working Group, Tech. Rep., November 2003. [Online]. Available: <http://www.w3c.or.kr/kr-office/TR/2003/ws-qos/>, last accessed Jan, 2011

[5] N. Dragoni, "Toward trustworthy web services - approaches, weaknesses and trust-by-contract framework," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, pp. 599-606, 2009.

[6] Z. Ying-Feng and S. Pei-Ji, "The model for consumer trust in C2C online auction," *ICMSE '06 International Conference on Management Science and Engineering*, pp. 125-129, Oct. 2006.

[7] M. N. Huhns and M. P. Singh, "Service-oriented computing: Key concepts and principles," *IEEE Internet Computing*, vol. 9, pp. 75-81, 2005.

[8] S. Kalepu, S. Krishnaswamy, and S. Loke, "Verity: a QoS metric for selecting web services and providers," *Proceedings Fourth WISEW*, pp. 131-139, Dec. 2003.

[9] Y. Wang and J. Vassileva, "A review on trust and reputation for web service selection," in *ICDCSW '07: Proceedings of the 27th International Conference on Distributed Computing Systems Workshops*. Washington, DC, USA: IEEE Computer Society, 2007, p. 25.

[10] E. Maximilien and M. Singh, "Toward autonomic web services trust and selection," in *ICSOC*, M. Aiello, M. Aoyama, F. Curbera, and M. P. Papazoglou, Eds. ACM, 2004, pp. 212-221.

[11] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decis. Support Syst.*, vol. 43, no. 2, pp. 618-644, 2007.

[12] Z. Malik and A. Bouguettaya, "Reputation bootstrapping for trust establishment among web services," *IEEE Internet Computing*, vol. 13, no. 1, pp. 40-47, 2009.

[13] T. A. Khopkar, "Provision, interpretation and effects of feedback in reputation systems," Ph.D. dissertation, School of Information, The University of Michigan, 2008.

[14] G. Zacharia, A. Moukas, and P. Maes, "Collaborative reputation mechanisms for electronic marketplaces," *Decision Support Systems*, vol. 29, no. 4, pp. 371-388, 2000.

[15] L. Zhengping, L. Xiaoli, W. Guoqing, Y. Min, and Z. Fan, "A formal framework for trust management of service-oriented systems," in *SOCA '07: Proceedings of the IEEE International Conference on Service-Oriented Computing and Applications*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 241-248.

[16] Y. Kim and D. Doh, "A trust type based model for managing QoS in web services composition," *International Conference on Convergence Information Technology*, vol. 0, pp. 438-443, 2007.

[17] L.-H. Vu, M. Hauswirth, and K. Aberer, "QoS-based service selection and ranking with trust and reputation management," vol. 3760 LNCS, Agia Napa, Cyprus, 2005, pp. 466-483.

[18] E. Maximilien and M. Singh, "Reputation and endorsement for web services," *SIGecom Exchanges*, vol. 3, no. 1, pp. 24-31, 2002.

[19] S. Jin-Dian, G. He-Qing, and G. Yin, "An adaptive trust model of web services," *Journal Wuhan University Journal of Natural Sciences*, 2005.

[20] 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, 1995.

[21] T. Kautonen and H. Karjaluo, Eds., *Trust and New Technologies: Marketing and Management on the Internet and Mobile Media*. Edward Elgar, 2008.

[22] S. Ran, "A model for web services discovery with QoS," *ACM SIGecom Exchanges*, vol. 4, no. 1, pp. 1-10, 2003.

[23] W. D. Yu, R. B. Radhakrishna, S. Pingali, and V. Kolluri, "Modeling the measurements of qoS requirements in web service systems," *Simulation*, vol. 83, no. 1, pp. 75-91, 2007.

[24] L. O'Brien, P. Merson, and L. Bass, "Quality attributes for service-oriented architectures," in *SDSOA '07: Proceedings of the International Workshop on Systems Development in SOA Environments*. Washington, DC, USA: IEEE Computer Society, 2007, p. 3.

[25] A. Moorsel, "Metrics for the internet age: Quality of experience and quality of business," 5th Performability Workshop, 2001.

[26] S. Lee and D. Shin, "Web service QoS in multi-domain," vol. 3, feb. 2008, pp. 1759-1762.

[27] W. Rahman and F. Meziane, "Challenges to describe QoS requirements for web services quality prediction to support web services interoperability in electronic commerce," in *Proceedings of the 10th IBIMA Conference on Innovation and Knowledge Management in Business Globalization, Kuala Lumpur, Malaysia, 30 June - 2 July 2008*, 4 (6), pp. 50-58.

[28] D. Menasce, "QoS issues in web services," *Internet Computing, IEEE*, vol. 6, no. 6, pp. 72-75, Nov/Dec 2002.

[29] D. Hoyle, *Automotive Quality Systems Handbook*, 2nd ed. Elsevier Ltd, 2005.

[30] K. Larson, "The role of service level agreements in it service delivery," *Information Management and Computer Security*, vol. 6, no. 3, pp. 128-32, 1998.

[31] T. Grandison and S. Sloman, "A survey of trust in Internet applications," *IEEE Communications Surveys and Tutorials*, vol. 3, no. 4, 2000.

Towards Peer Selection in a Semantically-Enriched Service Execution Framework with QoS Specifications

Jun Shen, Ghassan Beydoun
Of Faculty of Informatics
University of Wollongong
Wollongong, Australia
{jshen, beydoun}@uow.edu.au

Graham Low
Of School of Information Systems, Technology and
Management
University of New South Wales
Sydney, Australia
g.low@unsw.edu.au

Brian Henderson-Sellers, Shuai Yuan
Of Faculty of Engineering and Information Technology,
University of Technology, Sydney
Sydney, Australia
brian@it.uts.edu.au, shyuan@eng.uts.edu.au

Abstract— This paper promotes an ontology-based multi agent system (MAS) framework to facilitate Peer-to-Peer (P2P) service selection with multiple service properties. P2P-based service has emerged as an important new field in the distributed computing arena. It focuses on intensive service sharing, innovative applications and compositions, and, in some cases, high performance orientation. However, one of the remaining challenges for the P2P-based service composition process is how to effectively discover and select the most appropriate peers to execute the service applications when considering multiple properties of the requested services. By introducing an ontology, different ontology-based e-service profiles can be proposed to facilitate handling multiple properties and to enhance the service oriented process in order to achieve the total or partial automation of service discovery, selection and composition. In this paper, we present a conceptual framework for peer selection with a preliminary mathematical model and a selection process, so as to enhance the P2P-based service coordination system and its components.

Keywords- *semantic Web services; quality of service; WSMO; peer-to-peer.*

I. INTRODUCTION

With the increasing popularity and growth of Web services, many researchers have been interested in developing effective e-service or e-business applications based on various existing components for agent-based systems [1]. In a multi agent system (MAS) composed of a heterogeneous collection of agents with distinct knowledge-bases and capabilities, coordination and cooperation between agents facilitate the achievement of global goals that cannot be otherwise achieved by a single agent working in isolation [2]. The unique characteristics of a MAS have rendered most standard systems development methodologies inapplicable, leading to the development of Agent Oriented Software Engineering methodologies [3], [4].

However, along with a soaring number of Web services developed in agent-oriented decentralised environments, it is essential to consider the quality of service (QoS) for agents when running business processes. It is obvious that the dynamics and heterogeneity of distributed services become extremely important to both service requestors and service providers. Nevertheless, most research works presented so far are predominantly syntactic and have not truly incorporated semantic ontology approaches for service description and composition within a realistic business context. The discovery and integration of a new service into an existing infrastructure is yet to be fully automated and currently requires significant human effort. As a result, it is problematic that traditional methodologies cannot effectively and autonomously conduct service discovery and composition in a complex dynamic environment. Moreover, the QoS specifications proposed in the literature (e.g., [5], [6], [7]), are yet to agree on common defining concepts.

A set of non-functional properties in Web Service Modelling Ontology (WSMO) [8] ideally can be used as a discriminating factor to refine P2P-based Web services, so as to provide a more reliable service selection in business workflows. In this paper, we present a scalable WSMO-based conceptual framework to describe QoS and other features of Web services in a P2P-based environment. We also sketch an automatic concomitant semantic Web services selection process to automatically find appropriate Web services that effectively fulfil the requestor's requirements. Hence, we design an approach to select the most appropriate peers that will foster a better service composition according to semantics of the user's request.

The rest of this paper is structured as follows. Section II illustrates our P2P architecture approach. Section III presents our QoS model and WSMO integration in our P2P framework and sketches a practical solution for selecting appropriate peers with multiple properties, specified by our service quality conceptual model. Section IV is a discussion of other related work, while Section V concludes the paper.

II. P2P MAS TO COMPOSE SEMANTIC WEB SERVICES

Generally, in a P2P MAS architecture dedicated to sharing resources, the MAS acts as an interface to a set of resources distributed across a network. Every user of a resource has an agent acting on his/her behalf. This *user interface* agent seeks the resource that its master user requests. In addition, every resource, which a user of the system would like to share in return, would have a *resource keeper* agent that also belongs to the system and acts as a gate keeper to this local repository of resources that it shares with other peer *user interface* agents as they broadcast their requests. In this architecture, all agents co-operate fulfilling queries and having access to their repository of resources whenever a query received can be assisted by their local resources. Resources shared can be information (files of data, music, etc.) as specified in systems similar to Klampanos and Jose [9] and Mine et al. [10], or alternatively, they can be services as specified in this paper or in [11].

In our proposed P2P framework, the MAS consisting of all cooperating *user interface* and *resource keeper* agents respond to requests by a user (e.g., a service requester, a software developer, a human web user) represented by an *interface* agent in the P2P network that acts on their behalf. In our description of the system here, our focus is not on the agent oriented analysis of such a system, rather it is on the role of a quality of service ontology and domain ontologies in such a system to maintain the function of such a system. Therefore for the sake of simplicity and without loss of generality, we merge the roles of *user interface* and *resource keeper* as well as the potential role of *history manager* into one agent. This agent aims to fulfil the request, e.g., locates services and responds to queries by other similar agents. The collection of all these agents and agents assisting them in their tasks form a P2P community-based cooperative MAS. For composing services using their semantics, a P2P MAS is shown in Figure 1. An agent (an ellipse in Figure 1) representing a user (a hexagon in Figure 1) has access to a knowledge base containing services/resources that the user is willing to share with other users. Each service/file/resources (a cylinder in Figure 1) is identified by a unique identifier within the P2P network (e.g., Service identifier, HTML, pdf, music or video).

As agents automatically interact on behalf of users seeking services to be composed, communities of interest begin to emerge. These communities may overlap. Providers and users of services may belong to more than one community; for instance a service to ‘open an account’ may belong to the community of banking developers as well as that for insurance developers. As more and more services are composed, agents become more efficient and effective by interacting with the agents in the communities most likely to be able to provide them with service components. The P2P system is responsible for locating sites where candidate services are available, based on the previous

requests made. The mediation between service requesters and providers is always done by the system. When an agent makes a service request, a candidate agent provider responds either by providing details about services they can supply, or by refusing the service. When all responses are received, the requesting agent combines and refines the results to compose a list of services that can be composed to fulfill the request. The requester agent can then select services it wants to compose and initiates the composition.

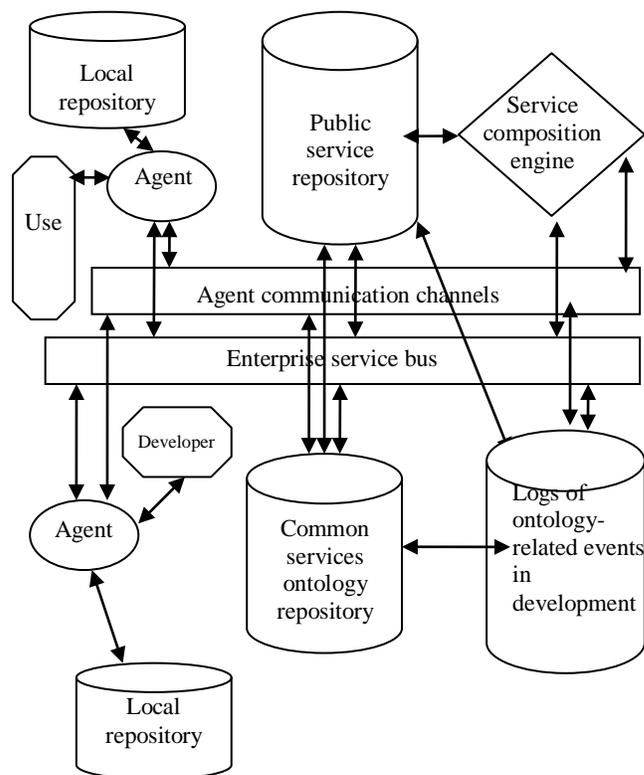


Figure 1. The P2P Multi agent system is the collection of the agent assistants and any supporting specialized agents.

After a successful composition a requester’s knowledge base is updated to now include the received and the composed services. Similarly for all agents involved in processing a service request, their knowledge base is also updated with additional information reflecting the domain and attributes of the requester agent. This information is used in future service requests. That is, as agents interact they develop awareness of the services possessed by their peers and which peers may be interested in the services that they themselves have. Each agent keeps a record of its history of service sharing in order to evaluate the quality of service (QoS) and to use this for future service requests. The collection of this history is in essence a distributed QoS ontology distributed across providers. The QoS ontology will provide assessments of past queries and providers. It is

used to produce short lists of candidate nodes for future queries, by calculating the similarity between the current query and a past query and its QoS. In a fully evolved P2P system, agents may use this QoS knowledge about other users' interests to request/negotiate for information from their peers when they do not know who has services of interest. New providers are constantly added to the history, expanding the user-agent's contact circle.

The proposed strategy of service sharing is domain independent and can be applied to any domain that can be prescribed by an ontology. With appropriate measure of the quality of services located by agents, the proposed P2P service execution system subsequently allows dynamic composition of Web services in a highly distributed and heterogeneous computing environment [11] that is adapted from [12] to highlight how ontologies can be used by taking advantage of semantically driven composition of services as is often advocated, e.g., in [13]. We aim to have the system provide, to both service requestors and service providers, the Quality of Service (QoS) evaluation. The system will identify the service providers' capability and performance so as to enhance the service composition for service clients over the real distributed service network. Due to the complexity of QoS metrics [1], [7], a well-defined QoS service description does not actually exist. With a P2P architecture, the QoS is gauged by a service client through cooperative interactions with other peers that can potentially provide the service. The scope of using ontology-based profiles in this MAS development is possible since most of the current work focuses on the definition of a QoS ontology, vocabulary or measurements, and, to a lesser extent, on a uniform evaluation of qualities. To provide a concrete measure to assess candidate services located by agents, we propose to exploit the Web Service Modelling Ontology (WSMO) [8] as a complementary conceptual framework to create the QoS ontology to describe various perspectives on Web services, thus facilitating the integration of the services. In Section III, we outline the components of the evaluation function, which can be enacted to a specified domain. Subsequently, we envisage Problem Solving Methods [11] as units of analysis corresponding to shared services that can be dynamically selected by agent communication sessions at runtime to best suit the service or the Quality of Service (QoS) required to match the requested service level agreement (SLA). This selection will be made using a Peer to Peer (P2P) searching mechanism to locate appropriate services from other peer agents. Cooperative communication between agents about their existing services, their past service requests (and who fulfilled them in the past) and their performance will enable service requestors to locate the peer service provider with the most suitable QoS.

III. PEER SELECTION WITH WSMO QoS CONCEPTUAL MODEL

First, WSMO defines four high-level notions that relate to semantic Web services, namely Ontologies, Goals, Mediators and Web services. *Ontologies*: are viewed as formal and explicit specifications of shared conceptualizations [8]. They define a common agreed-upon terminology by providing concepts and relationships among the set of concepts from a real world domain. *Goals*: are depictions of the expectations a service requestor may have when seeking for a service based on the following aspects: functionality, approach and quality of service. *Mediators*: coordinate the heterogeneity problem that occurs between descriptions at different levels [14]: data level - different terminologies, protocol level - different communication behaviour between services, and process level - different business processes. WSMO defines four types of mediators: OO Mediators connect and mediate heterogeneous ontologies, GG Mediators connect Goals, WG Mediators link Web services to Goals, and WW Mediators connect Web services resolving mismatches between them. *Web services*: are descriptions of services that are requested by service requestors, provided by service providers, and agreed between service providers and requestors.

Non-functional properties are usually utilised to describe non-functional aspects such as the creator and the creation date, and to provide natural-language descriptions, etc. All of the four WSMO elements have their own non-functional properties. In this paper, however, our QoS extension is of the same nature as the notion of non-functional properties in "*Web services*". In other words, we mainly introduce descriptors of QoS, such as performance, availability, spatial features of distributed services, etc. The incorporated QoS properties could also be used in parallel with existing non-functional attributes proposed by other WSMO elements.

We develop the non-functional properties in WSMO in order to support adaptive P2P-based service composition. These properties are domain-independent and can be used by agents assuming coordinator roles in our framework at runtime (as described in Section II). In using these properties, an emerging organisation of the peer/agent selection process and distribution of tasks is enacted at runtime. The resultant decentralised architecture is coordinated and self-managed effectively with services being allocated to peer/agent hosts, who are able to communicate with each other according to a real business process agreement or standard workflow definitions. In the rest of this paper, we present a more effective and qualitative representation to enable peers to evaluate candidate composition (in Section III.A) and select most appropriate peers (in Section III.B) for a requested service in a P2P information system.

A. A QoS Mode

The P2P-based service selection problem can be generally formulated as the following: Consider P as a set of composing agents, $P = \{p_1, p_2, \dots, p_N\}$, where each p_i ($i=1$ to N) is an agent that gets involved in the composition of a number of services from the set S covering M atomic services, $S = \{s_1, s_2, \dots, s_M\}$. Each atomic service (s_j) cannot be allocated to multiple peers, so let $x_{ij} = 1$ if atomic service (s_j) is allocated to Peer (p_i) and $x_{ij} = 0$ otherwise, and a Peer (p_i) can be allocated with a set of atomic services: $A_{p_i} = \{s_1, s_2, \dots, s_{m_i}\}$. Moreover, let $Q_{p_i, s_j} = \langle RT, CT, AV, RB \rangle$ denote the QoS features of Peer (p_i) for atomic service (s_j), and $\langle RT, CT, AV, RB \rangle$ represent ResponseTime, ComputationCost, Availability and Reliability. To optimally perform the service composition, the basic objective is to find a set of appropriate peers that makes response time and computation cost as small as possible, while keeping the availability and reliability as large as possible. Therefore, 4 sub-objectives can be defined as $\min(\sum RT)$, $\min(\sum CT)$, $\max(\prod AV)$ and $\max(\prod RB)$:

$$O_1 = \min\left(\sum_{p_i \in P} \sum_{k=1}^{m_i} Q_{p_i, s_j}(RT), s_k \in A_{p_i}\right) \quad (1)$$

$$O_2 = \min\left(\sum_{p_i \in P} \sum_{k=1}^{m_i} Q_{p_i, s_j}(CT), s_k \in A_{p_i}\right) \quad (2)$$

$$O_3 = \max\left(\prod_{p_i \in P} \prod_{k=1}^{m_i} Q_{p_i, s_j}(AV), s_k \in A_{p_i}\right) \quad (3)$$

$$O_4 = \max\left(\prod_{p_i \in P} \prod_{k=1}^{m_i} Q_{p_i, s_j}(RB), s_k \in A_{p_i}\right) \quad (4)$$

However, in order to consider the four objectives as a whole, it can be set as:

$$F = \max\left(\frac{w_3 \cdot O_3 + w_4 \cdot O_4}{w_1 \cdot O_1 + w_2 \cdot O_2}\right) \quad (5)$$

where $\{w_1, w_2, w_3, w_4\}$ denote the weights for the four QoS properties: RT , CT , AV and RB . Subject to the following constraints: $\sum_{i=1}^N x_{ij} = 1, j = 1, 2, \dots, M$,

$\sum_{j=1}^M x_{ij} = m_i, i = 1, 2, \dots, N$, $\sum_{i=1}^N m_i = M$ these respectively

ensure there is no conflict between peers to conduct atomic service allocation, and guarantee that the number of allocated atomic services of a peer are valid.

B. Selecting Peers with WSMO Enriched Non-functional Properties

Based on [7], we define an extensible class QoSProperty that aims to extend nonFunctionalProperties class in WSMO for P2P-based service selection.

```

Class nonFunctionalProperties other existing
properties...
    hasQoSProperty type QoSProperty
Class QoSProperty sub-Class
nonFunctionalProperties
    hasPropertyName type string
    hasPropertyValue type {int, float, long,
others}
    hasPreferredValueType type {low, high}
    hasWeight type float
    
```

Each QoS Property is generally described by *PropertyName* and *PropertyValue*. For the purpose of QoS-based selection, two additional properties are defined: *hasPreferredValueType* and *hasWeight*. The *hasPreferredValueType* property represents the desired trend. For example, the lower the response time is, the better QoS that could be achieved. The *hasWeight* is a value denoting the weight of the property, especially when synthetically measuring several different property metrics. In this context we define the weight value within the range [0, 1], while different end users may have different weight values for their service requirements.

For instance, a peer's "ResponseTime" can be described in Web service profiles as following:

```

dc _"http://purl.org/dc/elements/1.1#" ,
webService _http://example.org/ LoanApprove
nonFunctionalProperties
    dc#title hasValue "Peer 1"
    dc#description hasValue "ResponseTime for
LoanApprove process by peer 1"
    hasPropertyName hasValue _string
("ResponseTime")
    hasPropertyValue hasValue _int ("500")
    hasPreferredValueType hasValue _string
("low")
    hasWeight hasValue _float ("0.8")
endNonFunctionalProperties
    
```

In order to evaluate different non-functional properties of e-service peers, the important concepts in our modelling are: *PreferredValueType*, *Weight* and *Unified Value*. *PreferredValueType* has two possible values, "low" and "high". We utilise them to identify two types non-functional properties. For example, "ResponseTime" usually is expected to be as short as possible when choosing an appropriate peer, so the PreferredValueType of "ResponseTime" is "low". Likewise, "ComputationCost" also usually relates to "low", as no-one would prefer to choose a service with an expensive computation. However, "Reliability" and "Availability" often fit into the "high" category, since their values are often expected to be as high as possible. Accordingly, all peers' various properties can be categorised into the two types. With regard to "Weight", it indicates the importance and priority of certain properties during the service composition, so that

the weight value varies from service to service and from property to property. Lastly, “Unified Value” gives each peer’s overall quality measure, which can be used to assess each peer’s capability to meet the requirements of a requested service.

To enable the peers’ coordinating agent to intelligently select peers and plan a whole composition process, we sketch a selection process to assign the atomic services to appropriate peers within the service composition (Figure 2) that addresses the allocation method for multiple peer profile specifications and takes into account the above formulated objective.

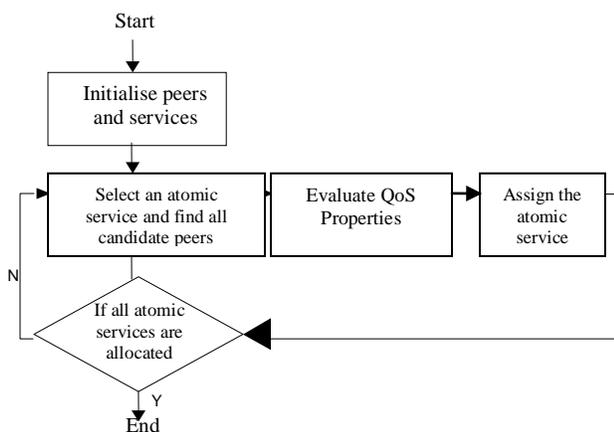


Figure 2. Allocating Atomic Services to Peers.

IV. RELATED WORK AND DISCUSSION

In recent years, the Semantic Web has become a hot topic and many researchers have turned their interests towards it. Functionality and non-functional properties are two essential aspects for semantic Web services. Functionality is used to measure whether this Web service meets all the functional requirements of an anticipated Web service, i.e. Web services matchmaking; while non-functional properties are qualified to evaluate the performance of the Web service. This has been viewed as a sufficient means to distinguish functionally similar Web services. In our previous work [15], [16], we presented a first sketch of the approach, although paying special attention to the extraction of the ontological description of services and the design of the selection process with OWL-S. The previous prototype was limited to a single specification, while it only considers “ResponseTime” as the selection criterion, which was not sufficient for effective services composition. Instead, this paper extends the description of non-functional properties via model-driven WSMO specification, and also presents an approach for the coordinator to automatically identify the best peers through unifying qualities and properties. In the rest of this section, we summarise and compare other approaches in this area. They are aimed at the same goal of

easing semantic Web services development for business process management systems.

Recently, QoS-aware service selection and composition have attracted considerable attention. Most related works focus on the development of QoS ontology languages and vocabularies, as well as the identification of various QoS metrics and their measurements with respect to semantic e-services. In [17], the authors have provided a QoS ontology as a complement for the DAML-S [18] ontology in order to provide a better QoS metrics model. Lee *et al.* [19] and Ran [6] emphasized a definition of QoS aspects and metrics, but have not included the extensible aspects in QoS, such as incorporating Geo features which we proposed in [15]. In [6], all of the possible quality requirements were introduced and divided into several categories, such as runtime-related, transaction support related, configuration management and cost related, and security-related QoS. Both of them present their definitions and possible determinants. Unfortunately, they are all too abstract to suit the implementation requirement. So, they did not tend to present a practical methodology for real service selection. In [5] and [7], the authors focused on the creation of QoS ontology models, which proposed QoS ontology frameworks aimed at formally describing arbitrary QoS parameters. Additionally, [20] and [21] attempted to conduct a proper evaluation framework and proposed QoS-based service selection, despite the authors failing to present a fair and effective evaluation algorithm. Furthermore, [22] and [23] also considered the evaluation of Quality of Service in the context of an overlay network or P2P principles. Based on [15], authors have evaluated our QoS solution, which was based on ACO and the ontology-based solution proposed in this paper. The evaluation results have been reported in [24].

V. CONCLUSION

In this paper, we have proposed a P2P-based service selection framework from the angle of an ontology-based P2P MAS. We described the operation of the P2P MAS and formulated the basic problem of service selection with multiple properties, and augmented the WSMO description by involving typical QoS perspectives. We have also designed a practical approach to facilitate peer selection. Our service peer selection model is expected to be enhanced in the near future by focusing on concrete and detailed geographic features for location-based services, and we will improve our framework for P2P-based workflow under more dynamic circumstances more effectively. Through this effort, we will be extending more complicated and useful specifications (e.g., representing realistic geographical knowledge) as well as protocols to enhance the accessibility, reliability and availability of e-services in decentralised information systems.

REFERENCES

- [1] A. Negri, A. Poggi, M. Tomaiuolo, and P. Turci, P., "Ontologies and web services: Agents for e-business applications." Proc. Fifth International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS '06), ACM Press, 2006, pp. 907-914.
- [2] M. Wooldridge, N. R. Jennings, and D. Kinny, "The Gaia Methodology for Agent-Oriented Analysis and Design," in *Autonomous Agents and Multi-Agent Systems*. The Netherlands: Kluwer Academic Publishers, 2000.
- [3] F. Bergenti, M. P. Gleizes, and E. Zambonelli (eds.), "Methodologies and Software Engineering for Agent Systems." The Agent-Oriented Software Engineering Handbook, MA: Kluwer Academic Publishers, 2004
- [4] B. Henderson-Sellers, and P. Giorgini (eds.), *Agent-Oriented Methodologies*, Idea Group, 2005, 413pp
- [5] I. V. Papaioannou, D. T. Tsesmetzis, I. G. Roussaki, I.G., and E. A. Miltiades, "QoS Ontology Language for Web-Services." Proc. 20th International Conference on Advanced Information Networking and Applications (AINA 2006), IEEE Press, 2006, pp.18-20.
- [6] S. Ran, "A model for Web services Discovery with QoS," ACM SIGecom Exchanges, vol. 4(1), 2003, pp. 1-10.
- [7] D. T. Tsesmetzis, I. G. Roussaki, I. V. Papaioannou, and M. E. Anagnostou, "QoS awareness support in Web-Service semantics," Proc. Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, 2006, pp. 128.
- [8] D. Roman, U. Keller, and H. Lausen, "Web Service Modeling Ontology," *Applied Ontology*, vol. 1(1), 2005, pp. 77-106.
- [9] I. A. Klampanos, and J. M. Jose, "An Architecture for Peer-to-Peer Information Retrieval," Proc. *SIGIR'03*, 2003, pp. 401-402.
- [10] T. Mine, D. Matsuno, A. Kogo, and M. Amamiya, "Design and Implementation of Agent Community Based Peer-to-Peer Information Retrieval Method," Proc. *Cooperative Information Agents (CIA2004)*, Springer, 2004, pp.31-46.
- [11] G. Beydoun, B. Henderson-Sellers, J. Shen, and G. Low, 2009, "Reflecting on Ontologies in Software Engineering: Towards Ontology-based Agent-oriented Methodologies," Proc. 5th Australasian Ontology Workshop (AOW 2009), Australia: ACS, 2009, pp. 23-32.
- [12] J. Shen, Y. Yang, and J. Yan, "A P2P based Service Flow System with Advanced Ontology-based Service Profiles," *Advanced Engineering Informatics*, vol. 21(2), 2007, pp. 221-229.
- [13] J. P. P. Sousa, E. Carrapatoso, B. Fonesca, M. G. C. Pimentel, and R. F. Bulcao-Neto, "Composition of Context-Aware Mobile Services Using a Semantic Context Model," *IARIA International Journal on Advances in Software*, vol. 2(2), 2009, pp. 1-13.
- [14] I. Toma, D. Foxvog, and M. C. Jaeger, "Modeling QoS characteristics in WSMO," Proc. Workshop on Middleware for Service Oriented Computing (MW4SOC'06), 2006, pp. 42-47.
- [15] S. Yuan, and J. Shen, "Mining E-Services in P2P-based Workflow Enactments", *Information Review*, Emerald Group Publishing, vol. 32 (2), 2008, pp. 163-178.
- [16] S. Yuan, and J. Shen, "QoS Aware Service Selection in P2P-Based Business Process Frameworks," Proc. 9th IEEE Conference on E-Commerce Technology and the 4th IEEE Conference on Enterprise Computing, 2007, Japan, pp. 675-682.
- [17] C. Zhou, L. T. Chia, and B. S. Lee, "DAML-QoS Ontology for Web services," Proc. International Conference on Web services (ICWS '04), 2004, pp. 472-479.
- [18] A. Ankolekar, M. Burstein, J. R. Hobbs, O. Lassila, D. L. Martin, D. McDermott, S. A. McIlraith, S. Narayanan, M. Paolucci, T. R. Payne, and K. Sycara, "DAML-S: Web Service Description for the Semantic Web" Proc. First International Semantic Web Conference (ISWC '02), 2002, pp. 23-30.
- [19] K. Lee, J. Jeon, W. Lee, S. Jeong, and S. Park, "QoS for Web services: Requirements and Possible Approaches," W3C Working Group Note 25, 2003. Available at: <http://www.w3c.or.kr/kr-office/TR/2003/ws-qos/>. Last accessed December 2010.
- [20] Y. T. Liu, A. H. H. Ngu, and L. Z. Zeng, "QoS computation and policing in dynamic Web service selection," Proc. International Conference on World Wide Web (WWW '04), 2004, pp. 66-73.
- [21] Y. Mou, J. Cao, S. S. Zhang and J. H. Zhang, "Interactive Web Service Choice-Making Based on Extended QoS Model," Proc. Fifth International Conference on Computer and Information Technology (CIT 2005), IEEE Press, 2005, pp. 1130-1134.
- [22] M. Song, and B. Mathieu, QSON, "QoS-Aware Service Overlay Network," Proc. Second International Conference on Communication and Networking in China (ChinaCom'07), IEEE Press, 2007, pp.739-746.
- [23] M. Kleis, K. Büttner, S. Elmoumouhi, G. Carle, and M. Salaun, CSP, "Cooperative Service Provisioning Using Peer-To-Peer Principles," Proc. First International Workshop on Self-Organizing Systems (IWSOS2007), Springer-Verlag, 2007, pp.73-87.
- [24] S. Yuan, J. Shen and A. Krishna, "Ant Inspired Scalable Peer Selection in Ontology-Based Service Composition", Proc. 9 International Workshop on Services Composition (SC-09), IEEE CS Press, 2009, pp. 95-102.

A View-based Approach for Service-Oriented Security Architecture Specification

Aleksander Dikanski, Sebastian Abeck
 Research Group Cooperation & Management (C&M)
 Karlsruhe Institute of Technology (KIT)
 Karlsruhe, Germany
 {a.dikanski, abeck}@kit.edu

Abstract—Developing secure software is still a software engineering challenge because of the complexity of software security. Yet integrating security engineering and software engineering is increasingly important, especially for service-oriented applications, as they are exposed to new security challenges due to their open nature. Current security engineering approaches do not consider existing security architectures, leading to redundant development of security artifacts. Further, present security architecture approaches do not provide relevant information to a security engineering process. Using a service-oriented and security architecture-centric approach for security engineering supports the development of secure service-oriented applications, as existing security solutions can be reused. In this paper, a model for service-oriented security architectures is presented, which provides apt information to different consumers, such as security engineering processes and business services, in the form of views to assist the consumers security goals. The architecture model is exemplified by specifying different views of a web service-based security architecture.

Keywords—security architecture; security engineering; service-orientation; web service, security services.

I. INTRODUCTION

Software engineering is focused on developing required functionality [1], but with increasing software support for business-critical processes, the importance of software security grows, in order to prevent financial loss, damage of reputation or data leakage [2]. Security engineering, i.e., the engineering of software which functions correctly under malicious attacks [1][3], requires the incorporation of security into all development process phases [4]. Yet the intrinsic specifics of security knowledge and the huge amount of complex security standards prevent such integration, leading to post-hoc consideration of security measures [5][6].

Due to the increasing use of networked software services, the focus of security engineering has shifted from software security to application security, i.e., using security products to secure already existing applications, components and services [7]. The paradigm of service-oriented architectures (SOA) established itself as the main architectural concept for today's enterprise information systems, as it allows traditional software systems to be restructured as reusable software services [8][9][10][11]. These software systems are now exposed to a vast quantity of new threats and attacks

they were not designed for and need to be protected by external security services [7].

Traditional security engineering approaches focus on eliciting and specifying security requirements and choosing appropriate security measures in order to secure single software systems. Current security engineering approaches for secure SOA applications tend to duplicate these procedures in order to provide security measures for single services. They thereby ignore existing security infrastructures, in which enterprises have invested large sums of money in to protect their existing software assets and which should be reused and restructured to security services according to the SOA paradigm. Additionally, existing security infrastructures provide constraints for security measures, which can be reused in a security engineering approach. On the other hand side, approaches for structuring the security infrastructure of an enterprise into a service-oriented security architectures do not consider this engineering view, instead they focus on the internal architecture and provided services without taking security requirements of service-oriented applications into account.

The contribution of this paper is the specification of a service-oriented security architecture model incorporating different interrelated views for security engineering, security infrastructure integration and security services in order to support the development and operation of secure service-oriented applications. The security engineering view provides development-time information artifacts, such as predefined security policy models and architectural constraints, as well as technology guidelines in order to support the development of secure service-oriented applications. The integration view specifies how existing security components and applications are restructured as security services. The service view specifies how application services can delegate their security requirements to the security services.

The rest of this paper is structured as follows: In Section 2, related approaches of security engineering and security architecture are discussed. In Section 3, our service-oriented security architecture model is presented and each view is discussed. Section 4 contains the specification of a concrete security architecture based on web service technology using our abstract specification, which was developed for a service-oriented navigation system. A conclusion and an outlook on future research directions conclude the body of this paper.

II. BACKGROUND AND RELATED WORK

A. Security Engineering

Software security knowledge reaches from secure code to enterprise wide security products such as firewalls. Each of these areas includes specific terms, standards and technologies that most developers are not familiar with. In order to support the development of secure software, security engineering intends to provide principles, methodologies, and tools for designing, developing, operating, and maintaining secure software systems [1].

Our research revealed only a few approaches for the methodical development of secure software. In [12][13], a general purpose security development process for traditional software is described using security patterns [15][16]. In [14], security problem frames, an adapted version of problem frames [17], are used to specify security requirements and to derive security components. Security engineering approaches for the development of SOA applications can be found in [18] and [6], which are motivated by the complexity of current web service security standards. Reference [18] uses patterns and model-driven software development techniques to automatically generate security standards artifacts. In [6] a process solely for the development of secure web service-based systems is presented. All of the approaches tend to generate new security measures and fail to map these to existing security infrastructures and services, which is a general requirement for SOA applications.

As the development of secure software needs to be embedded into a development method for software functionality, a larger amount of approaches focus on particular phases of security engineering processes. Misuse cases [19][20] and misuse activities [21][22] have been used to describe hostile intentions, attacks, and unwanted behavior in order to elicit security requirements. [23] uses extended problem frames to provide a formal specification of security requirements. Security patterns [15][16] are used to design security measures using best practice solutions. Each of these approaches supports a security engineering process, yet we have found no work, which shows their combined usage. It can also be argued, that these approaches are hard to adapt to the development of secure SOA applications, as it differs from the development of traditional software. Additionally, using each of these approaches leads to an increased complexity for developing secure measures, which matches the main development of the applications functionality.

In order to simplify the specifics of security, an increased interest in reusable security development artifacts can be registered in literature, such as reusable misuse cases, security requirements [24] and security patterns [15][16]. Yet developers still need to be supported in choosing appropriate items from a catalog of reusable security artifacts. A service-oriented approach can be helpful by providing a common set of reusable security artifacts to a development process, which are based on the currently used security measures implemented in a service-oriented security architecture, containing all information of previously developed secure SOA applications and services.

B. Service-Oriented Security Architectures

As there are several approaches for security architectures, we concentrated on the approaches for service-oriented security architectures in our literature review. In [25] the authors describe an event-driven reference architecture for three types of security services namely authentication, access control and identity management service, which communicate through the dispatching of security events. Similar service types are presented in [26], but the authors distinguish a service interface layer, providing security services to business-related services and layers of logical components implementing the security service functionalities. The logic components can be replaced by an existing security product as the authors show in [27]. The authors of [28] specify a technical security architecture for securing message exchange in business.

As each of these approaches focuses on different security measures, i.e., access control and message security, they never provide a complete picture of a security architecture. In neither of the approaches the application requirements, which lead to the specific structure of the architecture, are mentioned. As the services provided by a security architecture are determined by applications' security requirements, the presented architectures exhibit only very generic structures and need to be adapted to concrete security requirements.

The service-oriented integration of security products [29] is only directly considered in the approach of [26][27]. Their service access points only provide security services to service consumers, leaving the service implementation open. Using the service interfaces and event-driven approach for communication between the security services in [25] makes an integration not feasible as the efficient internal communication paths of security products are interrupted.

According to the SOA paradigm, services should provide their functionality through open and standardized interfaces in order to be reusable by different service consumers. Most of the aforementioned approaches rely on web service technology and related standards for service interface description. They thereby mix a technology-oriented view with the abstract description of their reference architectures, making it hard to adapt their approaches to other technology platforms. They also do not provide specific guidelines or conventions on how to apply the used standards, which would support the usage of the complex and flexible web service standards and related security standards.

Each approach has its advantages but neither alone provides enough support for software developers to handle software security complexity. So far, security engineering processes neglect existing and implemented security solutions, leading to the development of redundant security measures. Security architectures do not provide their intrinsic knowledge about implemented security solutions to security engineering processes. By providing such knowledge, software developers are supported in their task of choosing and implementing appropriate security measures.

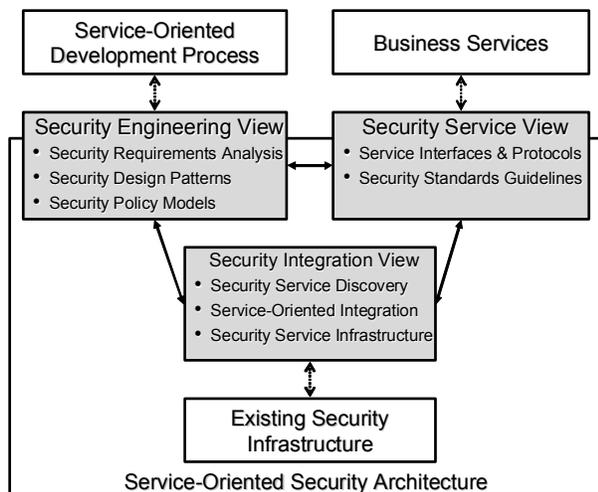


Figure 1. Service-oriented security architecture view model

III. SECURITY ARCHITECTURE VIEWS

In this section, a model for service-oriented security architecture specifications is introduced, which is arranged in views. Each view provides details about the security architecture and is tailored for a specific purpose, as it contains apt information for this purpose. We classify three views i) a security engineering view, which provides security information and supporting development artifacts for a software engineering process, ii) a service view, containing information about provided security functionality and usage protocols using the service-oriented paradigm as well as iii) an integration view, which specifies the details of integrating existing security products.

A. Security Engineering View

From a security engineering view, a security architecture provides relevant information to a service-oriented development process. The focus thereby lies upon service analysis and discovery as well as service design. Implementation and deployment phases become mostly obsolete, as existing and operated services are used to provide required functionality. Typically only specialized adapter components mapping proprietary interfaces of existing applications to service interfaces need to be developed [27]. In such cases, guidelines can be issued on how security standards are to be used to be compatible with the existing security architecture.

The goal of security analysis is to specify the security requirements of an application. An existing security architecture implements security requirements of previous service-oriented applications. Therefore information about previously identified threats, their corresponding attacks as well as the associated security requirements, implemented to mitigate or prevent the attacks, need to be documented and cataloged. In doing so, developers are supported in analyzing the security requirements of new applications. It would also complement approaches for catalogs of reusable security requirements including possible attacks and threats [24].

During a design phase, security requirements are mapped to appropriate security design. By security design we mean

static software structures such as security components and services as well as their dynamic interactions, which are added to the functional application design. A common approach for describing security design is through the use of security patterns, describing best practice solutions to reoccurring security problems. A security architecture should abstract from the implemented security measures, specify them using security patterns and set them into relation to previous security requirements, while keeping the implementation hidden to security service consumers.

Instead of directly attaching a security pattern to the application design, the patterns should be specified separately as part of the security architecture. Semantic annotations can be used to denote the location in the application design, at which security measures are to be adhered. This separates the security design from the application design, according to the separation of concerns paradigm. It also allows the security pattern specification to develop independently from the application design and for better reuse of the security design in different application designs.

Security policies play an important part in a security architecture, as they control the behavior of security components and services [25] and are thus a prime candidate for reuse. An example for security policies are access control policies of which several variations models for specific problems exists [30][31][32]. The security requirements of an application determine the kind of security policies the security architecture has to support. But reusing existing policy models of a security architecture, aides developers by providing a fixed and manageable set of policy models from which an appropriate model can be chosen from.

A security architecture might need to support multiple policy models, depending on different security requirements of different applications. For example, some applications might use a role-based access control policy [30], while others require a more fine-grained access control policy using attribute-based access control [32].

Predefined policies represent the standard security level of an organization, e.g., an enterprise wide constraint on integrity and confidentiality levels of message exchanged between application services. Issuing such policies, relieves developers from choosing an appropriate security policies and makes sure new application comply with the standard security level.

B. Security Service View

The security services view is concerned with the provided security services of a security architecture. It acts as a mediator between the abstract security engineering view and the low lever integration view. The abstract security patterns of the security engineering view are mapped to a specific set of security standards and technologies.

While a security architecture aims to centralize most of the security related functionality, there are security-oriented components which are hard to separate from the functional applications. A typical examples are policy enforcement points (PEP), which realize the result of policy decision points (PDP) at a resource [15]. The service view needs to

specify how such integration of security-related components and application-oriented services and components is performed. This is usually done using specific interactions protocols, which can be modeled using sequence diagrams of the Unified Modeling Language (UML, [33]).

A service-oriented economy requires interoperability of security services to exchange security information between business partners [34][35]. The usage of XML-based standards such as Security Assertion Markup Language (SAML), eXtensible Access Control Markup Language (XACML), WS-Security [36][37][38] is a good choice for interoperability with external partner but might have a negative effect on performance when used internally. The security services view therefore needs to provide specifications of the available security services, the use of related standards as well as proprietary standards or technologies both for internal use as well as for external use. Additionally the specific use of open standards needs to be documented, as they usually provide a large degree of flexibility.

C. Security Integration View

Concerning the integration view, the restructuring of existing proprietary security products to security services, by centralizing security components of existing applications is focused [26]. The other views build upon existing security products, as only their functionality can be provided as services and need to be abstracted from, in order to be used by other views. Due to its technical nature, this view is targeted at an organization's security experts.

We have shown how existing security functionality is integrated into a service-oriented security architecture by implementing web service adapters for a proprietary security product in [27]. Additionally we showed how standardized security policies are mapped to proprietary security product policies automatically. Due to place restrictions, we therefore will not focus on this view and refer to our previous work.

IV. A SERVICE-ORIENTED SECURITY ARCHITECTURE BASED ON WEB SERVICES

Based on the service-oriented security architecture view model, a concrete security architecture based on web service technology is described in this section. Due to place restrictions, we will concentrate on the access control functionality as well as on the security engineering and the service view, as we discussed the integration of security products in previous work [27]. We assume an existing security product, which provides access control evaluation functionality and which as been adapted to web service technology, if no such interface was provided by the security product itself.

Note also that we do not intend to describe a complete development process for secure service-oriented applications. Instead, we are focusing on single development artifacts that can be provided by examining and describing the security architecture according to the different views as described in the previous section.

A. Security Engineering Artifacts for Access Control

The security engineering view remains independent from concrete technologies and abstracts from the web service implementation of the security architecture, similar to the development process of the application services.

Service-oriented application development employs use cases to describe application services and detail their dynamics using process description languages such as the Business Process Modeling Notation (BPMN, [39]) [41]. The security architecture provides information about previously analyzed attacks and threats as well as the formulation of access control security requirements. Currently, misuse cases [19][20] and misuse activities [21][22] are used for this purpose, as they are regarded as initial step in determining security requirements [40].

A misuse case "Unauthorized Access" describes the intentions of an external attacker in accessing a resource, e.g., a service, which he is not allowed to access. The misuse case is stated more precisely through the definition of misuse activities, interacting with the business process of the service (Fig. 2(a)). The access control security requirement associated with the "Unauthorized Access" misuse case is expressed as an UML collaboration diagram (Fig. 2(b)) and can be instantiated using service candidates, which specify the parts of the process to be IT-supported.

The service candidates are the base for the development of the service design, i.e., provided service interfaces implemented by service components. Access control measures for services are provided using an access control service (ACS) and enforcement components, following the policy enforcement point (PEP) and policy decision point (PDP) pattern [15][16], to be reused in service design (Fig. 2(c)). The PEP's task is to query the PDP for access control decisions, which the latter calculates based on existing access control policies, and enforce the decision.

The access control policies used by the security architecture are based on the role-based access control (RBAC) model defined in [30] due to previous requirements (Fig. 2(d)). Developers are offered a common approach for specifying access control policies as well as previously defined roles and their associated entitlements, which can be reused in the development of new applications.

B. Access Control Service Specification

Security standards related to web services technology are used to describe the provided access control services. Building upon the service interface specification for the ACS provided in the security engineering view, following artifacts are provided: a web service interface for the access control service (WS-ACS) using the Web Service Description Markup Language (WSDL, [42]), an access control policy model using XACML [37] including policy specification guidelines, a query and response protocol for access control decisions using SAML [36].

As we have chosen to provide a RBAC model for SOA applications, we are using the RBAC profile for XACML [43] as the default specification for XACML access control policies.

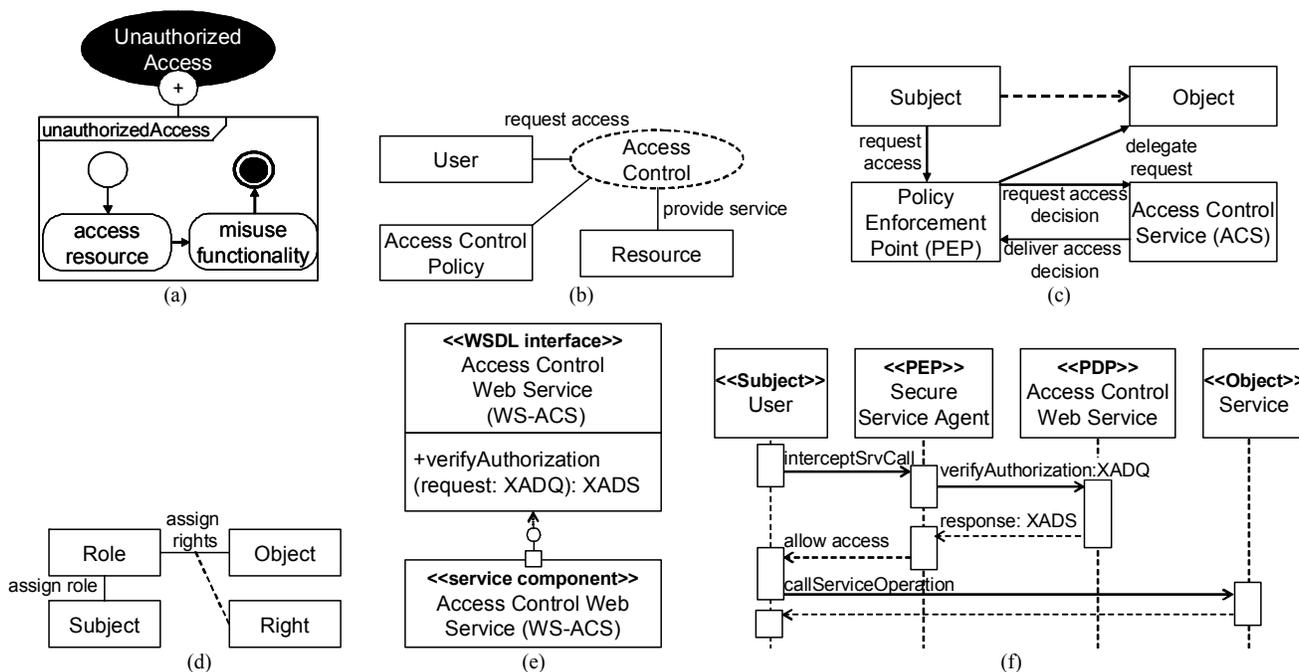


Figure 1. Views on an access control web service specification

While XACML provides access control decision query statements and response statements, it depends on supporting standards to implement a protocol and transport mechanism for them. For this purpose, SAML was chosen as a complement standard [44]. The data flow model of XACML is based on the PEP/PDP pattern discussed earlier. The PEP is implemented using the Secure Service Agent (SSA) pattern [45], which specifies a lightweight component for enforcing access control decisions on web services. The message interception is performed by the SSA by hooking into a web service frameworks’ message processing queue. The WS-ACS provides a WSDL interface containing an operation “verifyAuthorization”, accepting an XACML Authorization Decision Query (XADQ, [44]) message and returns a SAML statement (XACML Authorization Decision Statement, XADS, [44]) containing the result of the policy evaluation (Fig. 2 (e)).

On service invocation interception, the SSA sends an access control decision request to the WS-ACS using an XADQ request. The request includes the subject, which performs the access, the object, which is to be accessed as well as other contextual information needed to perform policy evaluation. On a positive outcome of the evaluation, the SSA allows the initial service invocation, while on a negative outcome the service access is prevented (Fig. 2 (f)).

V. CASE STUDY: DEVELOPING A SECURE CAMPUS NAVIGATION SYSTEM

We used our approach in the development of a secure service-oriented geographical information application for usage at the Karlsruhe Institute of Technology.

A. Case Study Scenario Description

The goal of the application is to support university students and lectures as well as employees and guests in

quickly finding and navigating to certain places or people on the campus. A basic functionality of the campus navigation application is to allow users to search for events, persons, rooms and buildings and to display them on a campus map as well as offering a routing service to the destination location.

The services needed for the implementation of the application are provided by different university organization units and are combined in order to create the application. Examples for provided services are a campus and building map service provided by the facility management unit, a staff information service provided by the human resources unit, which provides information about university employees and their corresponding location on the campus side, and event management services provided by different organization units, which provide information about upcoming events such as lectures, practical courses, conferences, etc. on the campus.

B. Approach to Specifying the Security Architecture

As the real-world scenario of the complete university security infrastructure is too complex to be explained in this paper, we will describe the development of the application prototype using the example security architecture described in the previous section. This conforms to the environment used for development and testing the application. We will also stick to the constraint of only discussing the access control aspect of the application.

In order to apply our approach, we first specified a security architecture for the application prototype by choosing existing security products, implementing the core security functionality. As a promising candidate for access control and authentication functionality we chose the open source “Identity Server” (IdS) product from WSO2 [REF], as it provided support for most of the relevant web service security standards.

As no previous knowledge for the product was available, the product's documentation was used to provide an abstract specification as required by our view model. The provided functionality, architecture and services were abstracted by matching architecture and security patterns already existent in literature. This process was mostly performed manually and often required deep product knowledge. In the real-world scenario the step was much easier to accomplish, as the required knowledge for the used product was provided by the security experts of the university.

As the IdS supported many security standards required for access control for web services the relevant security standards were used to describe the security service view. Additionally, due to this, it was not necessary to explicitly describe the security integration view.

We then proceeded to examine the existing literature for already described misuse cases, security patterns, etc. to specify the security engineering view of the architecture. We used a variant of the description of authorization misuse cases as provided by [19][20] as well as misuse activities for unauthorized access described in [21][22] to better match with the underlying IdS product. The IdS further implements a variant of the PEP/PDP access control pattern, using a central server as the PDP service component and an authorization module as a PEP component, which is placed in a service bus messaging and communication platform, utilized by the different web services. The choice for role-based access control was also determined by the security product and resembles the actual policy model used in the real-world scenario. The resulting description of the security architecture resembles the specification given in section IV.

C. Approach to Secure Development

The development process for the service-oriented application was mainly based on the method presented in [9], but any other approach could have been used as well, as our approach is process independent.

The development started with the requirements analysis of the application by defining use cases representing business services and stating their dependencies. As such a navigation use case was defined, which was related to user representation, staff information and event management use case. Additionally, the dynamic interactions of the use cases where modeled using BPMN processes.

A security analysis using the previously defined "Unauthorized Access" misuse case and activities (Fig. 2(a)) was performed, which resulted in the informal security policy, that students and unauthorized guests should not be able to search for the location of restricted rooms and areas upon campus, such as the storage facility of chemical substances of the chemistry faculty or the server rooms of the computing centre. A security requirement concerning access control for the search functionality was therefore specified.

The service design phase consisted of several service candidate specifications, which resulted from the use case and process descriptions. The service design was modeled using SoaML [REF] and included a service interface specification as well as a services architecture model, representing the dependencies between services. As several

services were already existent, their service interface representations were included in the service design.

As the access control service was predefined by the established security architecture, the access control requirement for the services was modeled by semantically annotating the services using an "authorization" marker tag. This left the service model mostly unchanged, allowing it to focus on its functional representation. The access control policies were described based on the RBAC policy model using a separately modeled UML class diagram. They were also attached to the corresponding services using semantic annotations

After the services were implemented using the provided web service framework of WSO2, the IdS server was configured with an XACML representation of the access control policy. The authorization module in the service bus was then able to determine the entitlements of user requesting access to the services of the application.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented a service-oriented security architecture view model, each of which provides the architectures' intrinsic security information for different purposes by reusing and complementing existing approaches. Three views were identified and described, a security engineering view, providing apt development artifacts, a security service view, describing provided security services, and a security integration view, for the integration and centralization of an organizations security infrastructure. We demonstrated our approach by specifying a security architecture using views and developing a secure service-oriented application using the intrinsic knowledge available by the security architecture.

In the future, we will further research how secure service-oriented applications can be developed using our security architecture view model. So far, we have neglected the issue of introducing new security measures into a security architecture in a methodical way, which is a topic we would like to research next. We also would like to introduce model-driven development techniques into our approach in order to automate the generation of security artifacts, such as security policies.

REFERENCES

- [1] R. J. Anderson, "Security Engineering: A Guide to Building Dependable Distributed Systems", John Wiley & Sons, 2001.
- [2] P.T. Devanbu and S. Stubblebine, "Software engineering for security: a roadmap", Proc. Intl. Conf. on the Future of Software Engineering (ICSE '00), ACM Press, 2000, pp. 227-239, doi: 10.1145/336512.336559.
- [3] G. McGraw, "Software security: building security in", Addison Wesley, Upper Saddle River, NJ, 2007.
- [4] G. McGraw, "Software security", IEEE Security & Privacy, vol. 2 (2), Mar./Apr. 2004, pp. 80-83, doi: 10.1109/MSECP.2004.1281254
- [5] A. Toval, J. Nicolás, B. Moros, and F. García, "Requirements reuse for improving information systems security: a practitioner's approach", Requirements Engineering, vol. 6 (4), 2002, pp. 205-219, doi: 10.1007/PL00010360.
- [6] C. Gutiérrez, E. Fernández-Medina, and M. Piattini, "PWSec: process for web services security", Proc. Intl. Conference on Web Services (ICWS '06), IEEE Press, Sept. 2006, pp. 213-222, doi: 10.1109/ICWS.2006.107.

- [7] J. Epstein, S. Matsumoto, and G. McGraw, "Software Security and SOA: Danger, Will Robinson!", *IEEE Security & Privacy*, vol. 4 (1), Feb. 2006, pp. 80-83, doi: 10.1109/MSP.2006.23.
- [8] D. Krafzig, K. Banke, and D. Slama, "Enterprise SOA – Service-Oriented Architecture Best Practices", The Coad Series, Pearson Education, 2005.
- [9] G. Engels et al., "Quasar Enterprise – Anwendungslandschaften serviceorientiert gestalten", Dpunkt Verlag, 2008.
- [10] N. Josuttis, "SOA in der Praxis – System-Design für verteilte Geschäftsprozesse", Dpunkt Verlag, 2006.
- [11] T. Erl, "Service-Oriented Architecture – Concepts, Technology, and Design", Pearson Education, 2006.
- [12] E. B. Fernandez, "A methodology for secure software design", Proc. Intl. Conference on Software Engineering Research and Practice (SERP'04), 2004, pp. 21–24.
- [13] E. B. Fernandez, M. M. Larrondo-Petrie, T. Sorgente, M. Vanhilst, "A Methodology to Develop Secure Systems Using Patterns", in Paolo Giorgini, „Integrating security and software engineering: advances and future visions”, IGI Press, 2007, pp 107-126.
- [14] D. Hatebur, M. Heisel, and H. Schmidt, "Analysis and Component-based Realization of Security Requirements", Proc. 3rd Intl. Conf. Third Intl. Conf. Availability, Reliability, and Security (ARES '08), IEEE Press, Mar. 2008, pp. 195-203, doi: 10.1109/ARES.2008.27.
- [15] M. Schumacher, E. B. Fernandez, D. Hybertson, F. Buschmann, P. Sommerlad, "Security Patterns : Integrating Security and Systems Engineering", Wiley series in software design patterns, Wiley, Chichester, 2006.
- [16] Ch. Steel, R. Nagappan, R. Lai, "Core Security Patterns: Best Practices and Strategies for J2EE(TM), Web Services, and Identity Management", Prentice Hall core series, Prentice Hall PTR, Upper Saddle River, N.J., 2006.
- [17] M. Jackson, "Problem Frames: Analyzing and structuring software development problems", Addison-Wesley, 2001.
- [18] N. A. Delessy and E. B. Fernandez, "A pattern-driven security process for SOA applications", Proc. 3rd Intl. Conference on Availability, Reliability and Security, IEEE press, Mar. 2008. pp. 416-421, doi: 10.1109/ARES.2008.89.
- [19] G. Sindre and A. L. Opdahl, "Eliciting security requirements with misuse cases", *Requirements Engineering*, vol. 10 (1), Jan. 2005, pp. 34-44, doi: 10.1007/s00766-004-0194-4.
- [20] I. Alexander, "Misuse Cases: Use Cases with Hostile Intent", *IEEE Software*, vol. 20 (1), Feb. 2003, pp 58-66, doi: 10.1109/MS.2003.1159030.
- [21] E. B. Fernandez, M. VanHilst, M. M. Larrondo-Petrie, and S. Huang, "Defining security requirements through misuse actions", in S. Ochoa and G.-C. Roman, "Advanced Software Engineering: Expanding the Frontiers of Software Technology", IFIP International Federation for Information Processing series, vol 219, Springer, Boston, 2006, pp. 123-137, doi: 10.1007/978-0-387-34831-5_10.
- [22] F. A. Braz, E. B. Fernandez, M. VanHilst, "Eliciting security requirements through misuse activities", Proc. 19th Intl. Conf. on Database and Expert Systems Application (DEXA '08), IEEE press, Sept. 2008, pp. 328-333, doi: 10.1109/DEXA.2008.101.
- [23] C. B. Haley, R. Laney, J. D. Moffett, and B. Nuseibeh, "Security requirements engineering: a framework for representation and analysis", *Transactions on Software Engineering*, IEEE press, vol. 34 (1), Jan./Feb. 2008, pp. 133-153, doi: 10.1109/TSE.2007.70754.
- [24] G. Sindre, D. G. Firesmith, and A. L. Opdahl, "A reuse-based approach to determining security requirements", Proc. 9th Intl. Works. on Requirements Engineering: Foundation for Software Quality (REFSQ '03), 2003, pp 16-17
- [25] E. Bertino and L. D. Martino, "A service-oriented approach to security - concepts and issues", Proc. 11th IEEE Intl. Works. on Future Trends of Distributed Computing Systems (FTDCS '07), IEEE press, Mar. 2007, pp. 31-40, doi: 10.1109/FTDCS.2007.6.
- [26] C. Emig, F. Brandt, S. Kreuzer, and S. Abeck, "Identity as a service - towards a service-oriented identity management architecture", in A. Pras and M. van Sinderen, "Dependable and Adaptable Networks and Services", Lecture Notes In Computer Science, vol. 4606, pp. 1-8, Springer, 2007.
- [27] A. Dikanski, C. Emig, and S. Abeck, "Integration of a security product in service-oriented architecture", Proc. third Intl. Conf. Emerging Security Information, Systems and Technologies (SECURWARE '09), IEEE press, June 2009, pp. 1-7, doi: 10.1109/SECURWARE.2009.8.
- [28] M. Hafner and R. Breu, "Security Engineering for Service-Oriented Architectures", Springer, Berlin, 2009.
- [29] D. Linthicum, "Next Generation Application Integration", Addison-Wesley Information Technology Series, 2004.
- [30] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models", *Computer*, vol. 29 (2), Feb. 1996, pp. 38-47, doi: 10.1109/2.485845.
- [31] M. A. Al-Kahtani and R. Sandhu, "A model for attribute-based user-role assignment", Proc. 18th Annual Computer Security Applications Conference, IEEE Press, 2002, pp. 353-262, doi: 10.1109/CSAC.2002.1176307.
- [32] E. Yuan and J. Tong, "Attributed based access control (ABAC) for web services", Proc. Intl. Conf. on Web Services (ICWS '05), IEEE press, July 2005, doi: 10.1109/ICWS.2005.25.
- [33] OMG, "Unified Modeling Language 2.3", 20 Nov. 2010; <http://www.omg.org/spec/UML/2.3/>, accessed 20 Nov. 2010.
- [34] P. J. Windley, "Digital Identity: unmasking identity management architecture (IMA)", O'Reilly, Beijing, 2005
- [35] C. Mezler-Andelberg, "Identity Management - eine Einführung: Grundlagen, Technik, wirtschaftlicher Nutzen", dpunkt Verlag, Heidelberg, Germany, 2008.
- [36] OASIS, "Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0"; <http://www.oasis-open.org/committees/download.php/35711/sstc-saml-core-errata-2.0-wd-06-diff.pdf>, accessed 20 Nov. 2010.
- [37] OASIS, "eXtensible Access Control Markup Language (XACML) Version 2.0", http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf, accessed 20 Nov. 2010.
- [38] OASIS, "Web Services Security: SOAP Message Security 1.1 (WS-Security 2004)", <http://www.oasis-open.org/committees/download.php/16790/wss-v1.1-spec-os-SOAPMessageSecurity.pdf>, accessed 20 Nov. 2010.
- [39] OMG, "Business Process Modeling Notation Version (BPMN) 2.0", <http://www.omg.org/cgi-bin/doc?dtc/10-06-04.pdf>, accessed 20 Nov. 2010.
- [40] I. A. Tøndel, M. G. Jaatun, and P.H. Meland, "Security requirements for the rest of us: a survey", *IEEE Software*, vol. 25 (1), Jan./Feb. 2008, pp. 20-27, doi: 10.1109/MS.2008.19.
- [41] M. Gebhart, J. Moßgraber, T. Usländer, and S. Abeck, "SoaML-basierter Entwurf eines dienstorientierten Überwachungssystems", Gesellschaft für Informatik (GI) Workshop zu SOA und Standardsoftware, Leipzig, Germany, Mar. 2010, pp. 360-367.
- [42] W3C, "Web Service Description Language (WSDL) 1.1", <http://www.w3.org/TR/wSDL>, accessed 20 Nov. 2010.
- [43] OASIS, "Core and hierarchical role based access control (RBAC) profile of XACML 2.0", http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-rbac-profile1-spec-os.pdf, accessed 20 Nov. 2010.
- [44] Organization for the Advancement of Structured Information Standards (OASIS), "SAML 2.0 profile for XACML", http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-saml-profile-spec-os.pdf, accessed 20 Nov. 2010.
- [45] C. Emig, H. Schandua, and S. Abeck, "SOA-aware authorization control", Proc. Intl. Conf. on Software Engineering Advances (ICSEA '06), IEEE Press, Oct. 2006, pp. 62-62, doi: 10.1109/ICSEA.2006.2613

End-user's Service Composition in Ubiquitous Computing using Smartspace Approach

M. Mohsin Saleemi
 Turku Centre for Computer Science (TUCS)
 Åbo Akademi University
 Turku, Finland
 msaleemi@abo.fi

Johan Lilius
 Department of Information Technologies
 Åbo Akademi University
 Turku, Finland
 johan.lilius@abo.fi

Abstract—This paper presents our architecture and overall process for creating end-user service compositions using smartspace approach. We have used OWL-S ontology language to describe the service capabilities semantically. We implemented the composition algorithm as the planning strategy for automatic service composition. This composition conforms to semantic graph-based techniques where atomic services are composed iteratively based on OWL-S service properties. We also presented a concrete example to show how this algorithm automatically discovers and composes services in a sequence that fulfills end-user's requests.

Keywords-smartspace; ubiquitous computing; composition.

I. INTRODUCTION

Recent advances in information and communication technologies have made available a wide range of devices and services to their users and hence making a device, service and information rich environment for them. It helps simplifying and managing our complex lives, e.g., the ubiquitous smartphone that manages our calendar, contacts and task list and helps us keep our lives organized, or the Personal Video Recorder (PVR) whose time-shift functionality allows us to watch TV programming when we want, not at the time prescribed by the broadcaster. However each of these devices is basically an island, with no proper connectivity between the applications. In order to take full advantage, devices need to interact with each other to perform different tasks. The problem thus is the closed and proprietary device architectures which have limitations in terms of scalability and interoperability. By exposing the internal data and functionality of the devices and ensuring interoperability of data, a whole new universe of applications will be possible. For example, your smartphone could notice that your favorite program will start in 5 minutes, based on your profile information or a fan page on facebook and the TV guide available on the broadcaster's web page. Then, it could use GPS to find that you are not at home, and deduces that it needs to start the PVR at home. Another example could be the composition of available services to form a complex composite service which is not otherwise possible.

To enable these kinds of cross-domain scenarios, there are many technical and conceptual problems to be solved. One way to address these issues is through the notion of a *smartspace*. A Smartspace is an abstraction of space that encapsulate both the information in a physical space as well as access to this information allowing devices to join and leave the space. In this way, smartspace becomes a dynamic environment whose membership changes over time when the set of entities interact with it to share information between them. For example, communication between the mobile phone and the PVR in the above scenario does not happen point-to-point but happens through the smartspace whose members are the mobile phone and the PVR. These device functionalities are available to the other elements of the members of the smartspace as services.

We have developed a prototype service composition architecture that supports service composition in smartspace environment. The system is able to achieve the desired user-tailored goals by composing the combination of available services in the smartspace. As part of our solution, we introduce a composition algorithm that find a set of candidate services which could be part of the composition. The complete realization is obtained by grounding of the selected services.

The rest of the paper is structured as follows. Section II describes the smart-m3 which is the underlying architecture of our work. In Section III, we presented the application development approach for smart-m3. This section also proposes our concept of *service* to the smart-m3 architecture and provides an example to illustrate the idea. In Section IV, we specify the service description language and the service composition. We then present our proposed system architecture of service composition in Section V. Section VI illustrates the implementation details of the example scenario of service composition using the language translation service. Section VII presents related research and we conclude the paper and give directions for the future work in Section VIII.

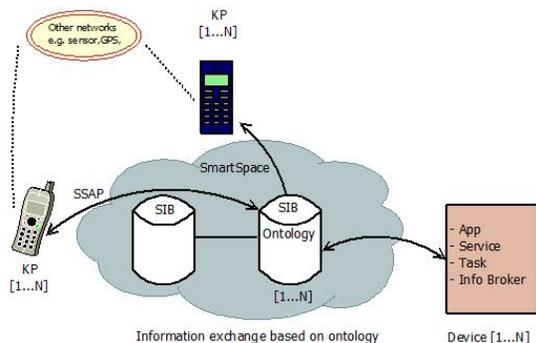


Figure 1. Smart-M3 Architecture

II. SMART-M3 ARCHITECTURE

The Smart-M3 architecture [9][2] provides a particular implementation of smartspace where the central repository of information is the Semantic Information Broker (SIB). The smart-M3 space is composed of one or more SIBs where information may be distributed over several SIBs for the later case. The set of SIBs in a M3 space are completely routable and the devices see the same information, hence it does not matter to which particular SIB in a M3 space a device is connected. The information is accessed and processed by the entities called Knowledge Processors (KPs). KPs interact with the M3 space by inserting, retrieving or querying the information in any of the participating SIBs using access methods defined by the Smart Space Access Protocol (SSAP). Smart-M3 provides information level interoperability to the objects and devices in the physical space by defining common information representation models such as the Resource Description Framework (RDF). In this way, it provides a device, vendor and domain independent solution for interoperability. Since smart-M3 does not constrain to a specific structure of information, it enables the use of ontologies to express the information and relations in an application. The ontology enables the KPs to access and process the information related to their functionality from the M3 space and hence it directs the KPs through the space. Figure 1 shows the overall Smart-M3 architecture.

III. APPLICATION DEVELOPMENT FOR SMART-M3

We chose the ontology-driven application development approach for smart-M3 and developed tools [7] for mapping of ontologies to Object Oriented Programming (OOP). Our approach consists of two parts. The first part is the generator that creates a static API from an OWL ontology. This mapping is done according to a set of static mappings. These mappings generate native Python classes, methods and variable declarations which can then be used by the KP developer to access the data in the SIB as structured and specified in the OWL ontology. The second part is the middleware layer which abstracts the communication with the SIB. Its main functionality to the generated API is triple

handling. This consists of inserting, removing and updating triples and committing changes to the smartspace. It also provides functionality for synchronous and asynchronous querying. Our approach enables application developers to use the generated API to develop new KPs and applications without worrying about the SIB interface as the generated API takes care of the connection to the SIB each time an object is created.

In this application development approach, the concept of application is not the traditional control-oriented application running on a single device but the application is constructed from a number of independently operated KPs which may run on different devices and group together to be perceived as a single application. For instance, chat, calendar synchronization and multiplayer games are examples of applications using this approach where a set of KPs each handling a single task run on multiple smart devices and coordinate and interact with each other through the SIB to make a complete application. This coordination between KPs are done in the form of data exchange through the SIB where KPs subscribe to or query for specific data to perform their specified task. Application ontologies are used to describe data in the SIB and directs the KPs to access and manipulate data related to their functionality.

We have taken the following approach to define our system.

The *Knowledge processor (KP)* is a single stateless software entity that reads or writes to the SIB either directly or by subscription. Each KP performs one piece of functionality. The functionality can have different forms and granularities. For example, we can identify device functionality which involves internal device resources such as a KP performing an MPEG encoding function within a PVR to convert analog signal to digital, or it can be a functionality related to computing of user data such as a KP translating a message from one language to another or a KP accessing calendar data in your smartphone. The KPs are accessible by other entities in the system only through the SIB.

The *Application* is the composition of behaviors of the KPs. The application exhibits the total functionality that is of interest to the user. We see applications as possible real-world scenarios that are activated by a particular set of KPs. For example, the scenario of the PVR and the smart phone described in the last section is an application "Record if I am not in home" using our approach. This application is built up by a number of KPs including a set of calendar KPs running in the smartphone and a set of PVR KPs running in the PVR to handle its tasks and functionality. The KPs running in the smartphone includes, for instance, a KP for reading calendar data, a KP for accessing profile information to check for favorite program, a KP for accessing the TV electronic program guide to check the schedule, and a KP for GPS data to find out the current location. The KPs running in the PVR includes, for instance, a KP to turn the PVR

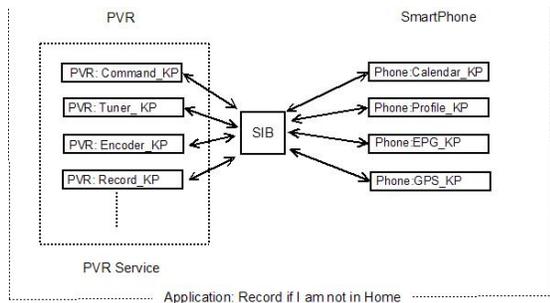


Figure 2. Application Scenario

on or off, a PVR tuner KP that receives the signals, an encoder KP that converts analog signal to digital, a KP to record digital stream to the internal hard drive of the PVR. These sets of KPs in the smartphone and the PVR coordinate by exchanging information through the smart-M3 space to enable this application. From the perspective of an application programmer, the PVR and the smartphone are two the entities in this application that need to be deployed. From deployment point of view, the PVR and the smartphone need to have the capability to tell the smart-space what is available and achievable. In order to make this declaration in a more common and well defined way, the idea of *service* comes in.

We propose to introduce the notion of *service* in Smart-M3 approach as a group of KPs that has a coherent set of functionality. We see a service as an interface to functionality that you can get through the smart-space. For example, the PVR in our example provides recording functionality and has a number of KPs each performing its specified function to implement the overall concept of TV recording. We see this as a PVR service which can be described by a service description language, e.g., OWL-S [1] and then be published in the smart-space to be available for the other KPs that are working within the application. They can call this service and add its functionality in the application without knowing how it actually works. The PVR service can be seen as a form of API where this API can have a number of other PVR functions as well which are not used in our scenario, e.g., you can use the audio-out jacks on the PVR to send the audio portion of programs and movies straight to the home stereo, which may produce better sound quality than the television. This functionality can be implemented by adding additional KPs. The coordination of functionality between the coherent set of KPs is done in the SIB. A single KP can always be seen as a service. The services can be stateless or state-full depending on their functionality. If a service is composed of only one KP that handle one specific task then it will be stateless. In our example application scenario, instead of having individual KPs for handling calendar activities in the application, a set of KPs dealing with calendar data group together to be perceived as a calendar service. These

calendar and PVR services are exploited by other KPs in this application, e.g., the KP handling GPS data to check whether it needs to activate the PVR to record a specific TV program. In this way, each service can act as a service provider, exposing its functionality to other KPs and services via the smart-space, and can act as a service requester, incorporating data and functionality from other services. These services are internal to the application, but can be reused by other applications after describing the properties and capabilities of the services using well defined service description languages.

TABLE I
CONCEPTS

Knowledge processor	Service	Application
Stateless entity	Stateless /stateful	Stateless /stateful
Handles one task	Composition of KPs	Composition of services and/or KPs
Read/write to the SIB	Described by description languages	Represents overall functionality
	Can be discovered and reused	

The SIB offers a persistent data storage back-end and is thought of as a dump and plain database that just gives access to the data. It does not provide any kind of services where some control structure and computations are involved. The coherent set of KPs provide the capabilities to enable services to the SIB in addition to just 'finding' the information. In this way, some computation and control structure is added as the services which have well defined interface and implemented by the set KPs using the object oriented programming approach. By storing the service descriptions in the SIB, it becomes possible to query the Smart-space for its Services. These services can be later discovered, invoked and reused by the other entities in the system. Hence we believe that the notion of service in this context provides advantages to the original smart-M3 approach.

IV. SERVICE COMPOSITION

The basic idea of composition is to use semantically related services in the system in such a way that their combination provides the desired goals which are not otherwise possible. For example, consider a scenario where a number of language translation services are available in a system each translating from one specific language to another. If the system receives a translate service request to a language that is not already available, it should be able to find the services that can combine to fulfill the request. This gives a broader view of service composition that also includes service discovery. The process of service composition is thus, to select a set of services which can possibly fulfill the request (Service discovery process)

and to compose these candidate services to form a single service. All of these tasks are done with the help of service specification languages that describe the services and enable automated or assisted searching of services that participate in the composition process. We propose to use OWL-S language for this purpose. There are several reasons for choosing OWL-S for service description. Firstly, OWL-S enables declarative advertisement of service properties and capabilities that can be used for automatic service discovery. Secondly, as we are using OWL ontologies for the domain concepts and application development for the Smart-M3, OWL-S describes the services in terms of capabilities based on OWL ontologies. Thirdly, OWL-S provides specification of prerequisites of individual services and a language for describing service compositions and data flow interactions. OWL-S can be used to construct complex composite services using OWL-S control constructs such as sequence, split, split-join, if-then-else, iterate, choice and loops. Currently we use only sequence construct in our application which requires a list of components to be done in an order where the output of the component A is compatible with the input of component B and so on.

The OWL-S provides three levels for service description: Service Profile, Process Model and Service Grounding. The Service Profile provides a general description for advertising, discovering and composition of the services. It includes both functional properties of services, IOPE (inputs, outputs, preconditions and effects) and nonfunctional service properties (name, text description, category and additional service parameters). The Process Model gives information about how a service performs its operation and describes the steps that should be done for the execution of the service. These steps include transformation of the set of inputs into the set of outputs and state transitions from one state to another when the service is carried out. The Service Grounding gives details about how to access a service using the specific message formats and platform provided protocols, for example, the Simple Object Access Protocol (SOAP) and HTTP used in accessing web services and the SSAP protocol in our case of the smartspace based applications. As the Service Profile describes the functional description that will be used in the service discovery and composition, we will only focus on Service Profile in this paper.

V. SYSTEM ARCHITECTURE

We have proposed a service composition framework for the smart-M3 approach. The architecture of the framework is presented in Figure 3. The goal is to enable both the end-users and the application designers to use the system for service discovery and composition in smart-M3 based application development. The functionalities supported by this framework are publishing the service descriptions in the SIB, searching the services relevant to a query or application

based on these descriptions, and composing the compatible services to form a single service to fulfill a given request. The framework consists of the following six layers.

1. User/Application Layer: The goal of this layer is to handle interaction with the end-users. It consists of user level application and the graphical user interface (GUI). The system is expected to receive the request from the end-users in the form-based query mechanism. It can then be converted to the OWL-S representation by the Interpretation layer of the system. The end-user interacts with the system using the application GUI that consists of a user query form. The end-user submits its request by filling the query form and specifies the desired goal in terms of its functional parameters. These parameters would be in the form of inputs, outputs, conditions and goals. This request enables the end-user to specify what he wants the service to do for him while abstracting the implementation of the service. For example, if a user wants to use the system for a language translation service, he submits the request in terms of the source language, the destination language, the goal and/or some precondition by using the end-user query form. The user will totally be unaware if the result of the request comes from a single service or composition of two or several services.

2. Interpretation layer: This layer is responsible for mapping the end-user's intent in some well defined notion, the OWL-S service ontology in our case. After the end-user specifies the desired goal using the application GUI in user/application layer, the user's request is converted into its equivalent form of the OWL-S service ontology by the interpreter agent in this layer. In this way, the end-users do not require to have extensive knowledge about the ontologies as the system handles their requests in natural language. The service descriptions in the request is then matched with the available services in the system to find out if the request can be fulfilled. This layer also contains a composition KP which is responsible for making the compositions of the available services if the request is not fulfilled by a single service. This composition KP relies on the Service Discovery component in the Semantic Service Layer. The Service Discovery component selects a set of services from the available services and passes this information to Service Composition which generates the composition by specifying the order in which each service is to execute to form a single composite service to fulfill the request. As the system is capable of interpreting OWL-S, the applications can also be given as OWL-S.

3. Semantic Service layer: The semantic Service layer handles the description, discovery and publishing of services. End-users can use the system to find out if it is able to do a particular thing such as translating from one language to another. There is another broader way in which services are written in OWL-S and need to be interpreted and executed by the system. For example, the PVR service and

the calendar service are provided by their service providers in OWL-S and the end-user application 'Record if I am not in home' interpret and execute these predefined OWL-S services in the system. As described in Section 3, a coherent set of KPs makes a service, e.g., the PVR service includes several KPs each performing a single task. Moreover, each KP handling a single task can also be seen as a service, e.g., each KP translating from one language to another is regarded as a separate service. In order to make these services available to the other entities in the system and to facilitate service composition, each of these services must be described using OWL-S. The semantic service layer is responsible for these service descriptions. After describing each service, the services need to be published. The service provider describes service functionality and other information which must be converted in OWL-S in order to store and publish in the SIB. The Service Discovery is done after services have been described and published and return a list of services as a result to the composition KP.

4. KP Layer: The KP layer is the low level layer which is responsible for bindings of the KPs. Each KP handles one task or functionality and interact with the repository to read and write required data. An end-user application contains a series of actions that need to be performed to fulfill its desired goal. These actions are actually implemented by the KPs in the system.

5. Middleware Layer: The middleware layer contains all the necessary components that are required to access the devices and services in the smartspace. It includes DIEM Mediator that provides a smartspace independent interface for accessing certain functionality of the SIB. This interface encapsulates the communication of KPs with the SIB and provides modularization. It provides the functionality of triple handling and synchronous and asynchronous querying. The middleware layer may optionally include other 3rd party middleware solutions such as UPnP middleware.

6. System Layer: The system layer consists of the SIB which acts as a persistent data storage for the information in the form of RDF triples. This layer also contains the operating system and other local device resources and provides a native interface to the upper layers.

In this system we are proposing the OWL-S language as a kind of interoperability format for KP-based smartspace applications. The notion of service in this kind of applications gives a well defined interface which provides benefits over traditional Smart-M3 application development approach which is based on only KPs. These benefits include, for instance, description, discovery and composition in a more structured way.

By stroing the service descriptions in the SIB, it becomes possible to query the smartspace for its services. From a programming point of view, the notion of services is a structuring mechanism that groups together several different KPs to abstract individual calls from the API of a device

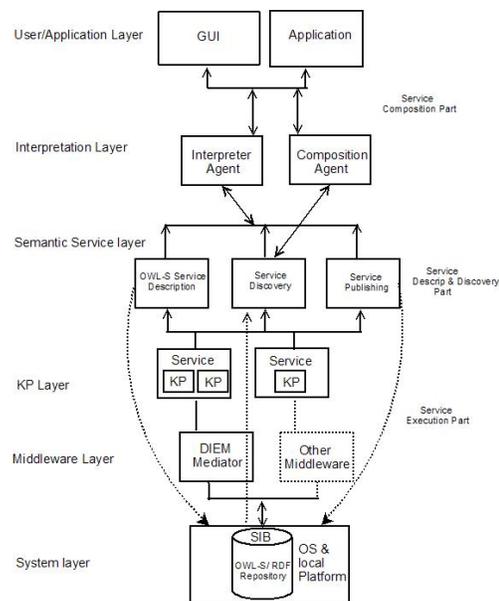


Figure 3. Structure of the System

that is present in the Smartspace. We make a dedicated space in the SIB, that can be used for registering the services. In this case the SIB will be the central part and all the requests and responses of the services will go through the SIB. For example, when a service requester submits a query to the SIB, the SIB will interpret the query and determine the match using the semantic description of services. After selecting the matching service it will construct a query to the service provider agent. The service provider agent interprets this query from the SIB, performs the service execution and generates a response to the SIB. The SIB interprets this response, transforms it to the response understood by the requester and sends it to the service requester.

VI. EXAMPLE SCENARIO: LANGUAGE TRANSLATION

To illustrate the approach, we consider the following service composition scenario where many KPs in the system coordinate each other to fulfill the end-user's request that is not possible with a single KP. The scenario is the language translation function which translate from one source language to the target language. Suppose we have many KPs in the system where each is capable of translating one particular language to another. We call each of the translate functions as a Service. An end-user uses the system and sends the query to translate a particular message in a source language, for instance, Finnish to a target language, for instance, Urdu. Since each KP has to perform its own dedicated single task, there is no KP that directly satisfies this request. In other words, there is no single service in the system that can perform this task. In this case, the system is able to find appropriate composition of different

services to satisfy the user's request. The system performs this composition automatically hence the end-user does not need to worry about invoking each individual service that takes part in the composite service.

A. Composition Algorithm

We have implemented a composition algorithm for service composition. The algorithm is implemented in python as we have only a python version of the SIB interface available at the time of writing this paper. The algorithm follows the Breadth First search pattern to find all the candidate services that can be included in the composition. The Breadth First search gives better results as it moves horizontally across each search path and finds the solution with the fewest possible services and does not face looping problem. As the SIB can read and write only RDF data, our algorithm use the same format to interact with the SIB to insert and query the RDF triples. The algorithm works in two parts. In the first part, the services which can be included in the composition are separated. In the second part, the composition is constructed based on the selected services using the approach that begins with inputs and preconditions and works forward in the direction of outputs and effects. This shows that each and every service should be described semantically in terms of its inputs, outputs, preconditions and possible effects in order to make this algorithm work. Our algorithm always finds a composition solution if it is available unless there are infinite services. If the composition is not possible from the available services in the system, it returns the message that the user's request cannot be fulfilled. The proposed algorithm can be evaluated based on its completeness and optimality. The time and space complexity depends on the particular service compositions. As our focus in this project are the smart devices, we have implemented the algorithm in a way that reduces the space complexity by ceasing the algorithm at some depth to avoid memory overflow. However, there will be compromises in terms of completeness because the algorithm may not find the solution within that search path.

When an end-user requests a specific service by making a query, the algorithm starts with checking the inputs and preconditions of the requested service and matching these parameters with the available services. All the services which have similar inputs and precondition parameters as the requested service may be possible candidates for service composition and are selected. The algorithm then works by constructing a graph of the services in a forward chain pattern by checking the outputs and effects of the previously selected services. This process continues until it reaches to the service having desired output. This service is placed at the leaf node of the graph and the composition is returned by traversing the graph.

B. Example

We use the example of language translation to illustrate the applicability of the proposed algorithm. We assume that the system has a range of different services provided by individual KPs where each KP performs only one task. An end-user wants to use the system to send the meeting invitations to all members of the project in their mother tongue. The original invitation is in the Finnish Language. The user sends the request to the system by using some GUI to translate the particular Finnish text into the desired language, for example Urdu. Suppose that there are many different KPs, each providing a translation service from one particular language to another. The Service handler takes this request and prepares the semantic description of this request in terms of inputs, outputs, preconditions and effects. In general, Inputs and Outputs are subclasses of a general class Parameter in the OWL-S service ontology. Every parameter has a type that can be specified using a URI to uniquely identify it. The type can either be a Class or a Datatype such as a number, a string etc. The Preconditions and effects represent more specific functional properties to easily discover the services. For example, assume the service request is to translate from a source language to a particular target language with the input and the requested output of type string. There might be services in the system that have the same input/output type but their goals and preconditions are different, such as a dictionary service which has precondition of the same input and output language.

This information is derived from the service profiles of each of the available services. The profile gives the details about each service. The following OWL-S statements shows the general profile of the language translation service.

```
<profile:Profile rdf:ID="Lang">
  <profile:serviceName
    rdf:datatype="http://www.w3.org/...#string">
  Lang Trans
  </profile:serviceName>
  <profile:hasPrecondition
    rdf:resource="#Lang_precond"/>
  <profile:hasInput>
  <process:Input rdf:ID="message">
  <process:parameterType rdf:datatype=
    "http://www.w3.org/...#anyURI">
  http://...#Message
  </process:parameterType>
  </process:Input>
  </profile:hasInput>
  ...
</profile:Profile>
```

These parameters are then matched with the available services in the system to find out if there is any single service that matches all these parameters. If there is any, a message is sent to its respective KP to execute the service using the text as input and the results are returned to the end-user. If there is no such match found, then the composition algorithm starts with finding all the services which has

the same input, preconditions and effects as the requested service i.e the services accepting Finnish language as input, having different input and output languages as preconditions and translation as the goal/effect. These services could be the possible candidates for the composition and hence separated by constructing a graph with each node representing a service in the same horizontal level. Suppose that the system finds the following services in this step.

- S1: <Input:Fin><Output:Spa><Effect:translation>
- S2: <Input:Fin><Output:Eng><Effect:translation>
- S3: <Input:Fin><Output:Ger><Effect:translation>

The algorithm constructs the graph by placing 'Fin' at the root and Swe, Eng and Ger as its immediate children. It then checks each immediate children by checking their outputs and effects and putting each matching services under their respective nodes. Suppose the following services are retrieved in this step.

- S4: <Input:Spa><Output:Eng><Effect:translation>
- S5: <Input:Ger><Output:Dek><Effect:translation>
- S6: <Input:Eng><Output:Urd><Effect:translation>
- S7: <Input:Eng><Output:Fre><Effect:translation>
- S8: <Input:Eng><Output:Swe><Effect:translation>
- S9: <Input:Ger><Output:spa><Effect:translation>

The algorithm then checks if the composition is possible from the selected services as the original requested output (Output: Urd) is found in this step. It traverses the graph to reach the root (Input: Fin) and finds the path. The resulting composition S2->S6 is returned to the service handler. It is important to note that based on the selected services, there is another composition possible for the same request <Input:Fin ><Output:Urd >which is S1->S4->S6. The algorithm always returns the composition which involves fewer services. If the service composition is not found in this step, the algorithm continues working until it reaches the desired output or there is no composition possible from the available services. For the later case, the algorithm returns a message of 'composition not possible' to the service handler. The algorithm can run indefinitely if there is very large number of services and it does not find any suitable composition to fulfill the request. This means that we need to apply some end point to limit the number of times it executes to reduce the memory and bandwidth requirements of smart devices. Figure 4 represents the results of the service composition using our algorithm.

The composite service produced by the composition of atomic services can then be described semantically so that it can be discovered or take part in other compositions later. This way, we are able to provide services that are not actually included in the system.

C. Service Grounding and Execution

After the service composition is created, the next step is to execute these services in order to get the desired goals. The service grounding prescribes the details of how to access

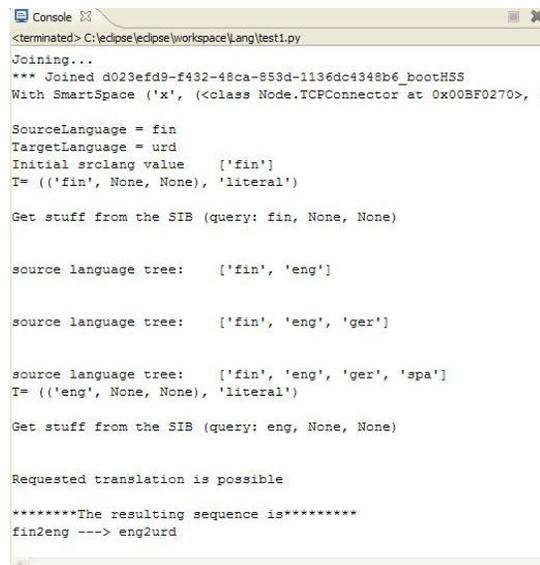


Figure 4. Composition results

the services in the composition such as message formats, protocols and invocation methods etc. used to interact with the composed services. It specifies a link between semantic and non-semantic description elements of the services. The grounding can include some or all of the following elements

Non-semantic elements:

- ServiceName -> Name of the service
- ServiceType -> Type of the service
- KPInfo -> Information of KP providing service

Semantic elements:

- ServiceInput -> Input of the service
- ServiceOutput -> Output of the service
- InputType -> Type of the Input parameter
- OutputType -> Type of the output parameter
- Preconditions -> Conditions met before execution
- ServiceEffects -> goal of the service
- MessageFormat -> Format acceptable by KPs
- Protocols -> protocols to interact with the SIB
- ServiceOntology -> specifies relations

As in our system implementation, every co-ordination between the KPs goes through the SIB, the execution of the services in the composition is done by invoking each individual service and passing data between the services in the order specified by the composition using the grounding elements described above. This invocation is accomplished by sending the messages to the service provider agents in their acceptable formats. This approach can undergo scalability problems in case of an excessive message and data passing between a large number of services in the composition such as in sensor applications.

VII. RELATED WORK

There are different approaches and architectures that address the issue of service composition. These approaches can be classified using several service composition features such as automatic composition [8], semi-automated composition [10], end-user interaction [11] and service specification language [6] etc. In [3], the authors give a comparison of different service composition approaches. A middleware solution for end-user application composition is provided in [5]. Other approaches of flexible service composition in mobile environments are described in [4][12]. While existing research efforts deal with these issues separately, there has been very limited work in ubiquitous service compositions in smart environment. In [13], the authors proposed a system consisting of a middleware and user-level tools that enable the end-users to combine, configure and control the services using their smart home devices. Each home device has interfaces and the end-user selects some devices and the system generates a set of compositions of selected devices in a way that these devices can collaborate to generate applications by using the domain knowledge and user inputs. This composition may result in an arrangement that has no meanings. They have used the Depth First Search (DFS) algorithm for generating the composition. The drawback with this algorithm is that it does not know which composition is better and which composition does not make sense. In our approach, each service is described semantically and this service description is stored in the SIB. It gives the advantage of query and accessibility of the services from the smart space.

VIII. CONCLUSION

In this paper, we expressed our ideas for providing services to the smart-m3 approach. These services are provided by means of sets of Knowledge Processors. We presented the system architecture and example services to illustrate our approach. A service composition algorithm is also presented with a concrete example. For future work, we are aiming to implement an efficient algorithm for discovering the services in the big search space of smart-m3. Furthermore, we need to use the ontology hierarchy to restrict the set of services considered for matching.

ACKNOWLEDGMENT

The research work presented in this paper is based on the ICT-SHOCK DIEM (Devices and Interoperability Ecosystem) project and the authors would like to acknowledge all the partners of this project.

REFERENCES

- [1] Owl-s services: <http://www.daml.org/services/owl-s/>. Accessed: October 2010.

- [2] Smart-m3 software at sourceforge.net, release 0.9.4beta, may 2010. [online]: <http://sourceforge.net/projects/smart-m3/>. Accessed: October 2010.
- [3] J. Bronsted, K. M. Hansen, and M. Ingstrup. A survey of service composition mechanisms in ubiquitous computing. In *Ubicomp 2007*, pages = 87-92.
- [4] D. Chakraborty, A. Joshi, T. Finin, and Y. Yesha. Service composition for mobile environments. *Mob. Netw. Appl.*, 10:435–451, August 2005.
- [5] O. Davidyuk, N. Georgantas, V. Issarny, and J. Riekk. MEDUSA: Middleware for End-User Composition of Ubiquitous Applications. In *Handbook of Research on Ambient Intelligence and Smart Environments: Trends and Perspectives*. IGI Global, 2010.
- [6] J. Dong, Y. Sun, S. Yang, and K. Zhang. Dynamic web service composition based on owl-s. *Science in China Series F: Information Sciences*, 49:843–863, 2006. 10.1007/s11432-006-2026-2.
- [7] A. Kaustell, M. M. Saleemi, T. Rosqvist, J. Jokiniemi, J. Liljus, and I. Porres. Framework for smart space application development. In *Proceedings of the International Workshop on Semantic Interoperability, IWSI*, 2011.
- [8] S. Majithia, D. W.Walker, and W.A.Gray. Automated web service composition using semantic web technologies. In *Proceedings of the International Conference on Autonomic Computing (ICAC04)*.
- [9] I. Oliver and J. Honkola. Personal semantic web through a space based computing environment. In *Proceedings of the 2nd IEEE International Conference on Semantic Computing*, 2008.
- [10] E. Sirin, J. Hendler, and B. Parsia. Semi-automatic composition of web services using semantic descriptions. In *In Web Services: Modeling, Architecture and Infrastructure workshop in ICEIS 2003*, pages 17–24, 2002.
- [11] Z. Song, Y. Labrou, and R. Masuoka. Dynamic service discovery and management in task computing. In *Mobile and Ubiquitous Systems: Networking and Services, MOBIQUITOUS 2004.*, 2004.
- [12] M. Valle, F. Ramparany, and L. Vercouter. Flexible composition of smart device services. In *The 2005 International Conference on Pervasive Systems and Computing(PSC-05), Las Vegas*, pages 27–30, 2005.
- [13] P. Wisner and D. N. Kalofons. A framework for end-user programming of smart homes using mobile devices. In *Proceedings of the Consumer Communications and Networking Conference, CCNC*, 2007.

Analysis and Verification of Web Services Resource Framework (WSRF) Specifications Using Timed Automata

José A. Mateo, Valentín Valero, Enrique Martínez and Gregorio Díaz

Department of Computer Science

University of Castilla-La Mancha

Albacete, Spain

{*jmateo, gregorio, valentin, emartinez*}@*dsi.uclm.es*

Abstract—Throughout the history of computing, engineers have used various formal methods to improve the quality of software and hardware. The next natural step is trying to exploit their advantages in the so-called new era of computing: Cloud Computing. In this paper, we present a first approximation about how to simulate and check the behaviour of these systems using timed automata through the model checking tool UPPAAL. We use Web Services Resource Framework (WSRF) as a standard intended to the modelling of distributed resources using Web services, and we apply formal techniques to WSRF specifications in order to analyse and verify these specifications.

Keywords-Web Services Resource Framework (WSRF); model checking; timed automata.

I. INTRODUCTION

The architecture that represents Web services has been widely accepted as a means of structuring the interactions between services in a distributed system. Nowadays, developers require more standardization to facilitate additional interoperability between these services. In January of 2004, several members of the organization *Globus Alliance* and the multinational company *IBM*, with the help of experts from companies such as *HP*, *SAP*, *Akamai*, etc., defined the basic architecture and the initial specification documents of a new standard for that purpose [6]. Web services Resource Framework (WSRF) has been inspired by the work previously done by *Global Grid Forum's Open Grid Services Infrastructure (OGSI) Working Group* [12]. Although a Web service definition does not consider the notion of state, interfaces frequently provide the user with the ability to access and manipulate states, i.e., data values that persist across, and evolve as a result of Web service interactions. However, the notion of stateful resources defined by the Web service implementation is not explicit in the interface definition. The messages that the services send and receive imply (or encourage programmers to infer) the existence of an associated stateful resource type. It is then desirable to define Web service conventions to enable the discovery of, introspection on, and interaction with stateful resources in standard and interoperable ways [4]. These observations motivated the WSRF approach to model Web services re-

source states. A WS-Resource is defined as the composition of a Web service and a stateful resource. WSRF allows WS-Resources to be declared, created, accessed, monitored for change, and destroyed via conventional mechanisms. WSRF consists of a set of five technical specifications that define the normative description of the WS-Resource approach in terms of specific message exchanges and related XML definitions.

In this paper, we propose the use of formal techniques and, more specifically, timed automata as a way to analyse and verify WSRF specifications. Thus, formal methods are used to write specifications that show the behaviour of the systems in a formal manner, and serve as the basis for system analysis to search for inconsistencies or errors in an early state of the development process, that is, before implementation. Furthermore, within formal methods environment, we use model checking techniques. Model checking [3] is an automatic technique for verifying finite-state reactive systems. In this approach, the specifications are expressed in a propositional temporal logic, and the reactive system is modelled as a state-transition graph (automaton). An efficient search procedure is used to determine automatically if the specifications are satisfied by the automaton. Model checking has a number of advantages over verification techniques based on automated theorem proving. The most important is that the procedure is highly automatic so it makes the testing phase faster. Typically, the user provides a high level representation of the model and the specification to be checked. The model checker will either terminate with the true answer, indicating that the model satisfies the specification, or giving a counterexample that shows why the formula is not satisfied.

As far as we know the literature in this field, no one has modelled the communication model in WSRF. Nevertheless, there are some works that use WSRF in a practical way. In [10], the authors presented a meta-model of a medical system for deriving clinical trial information management systems for collaborative cancer research across multiple institutions. With this meta-model, they extract the corresponding semantics in the Z formal specification language and the WSRF implementation in the real environment CancerGrid. The main difference with our work is the different

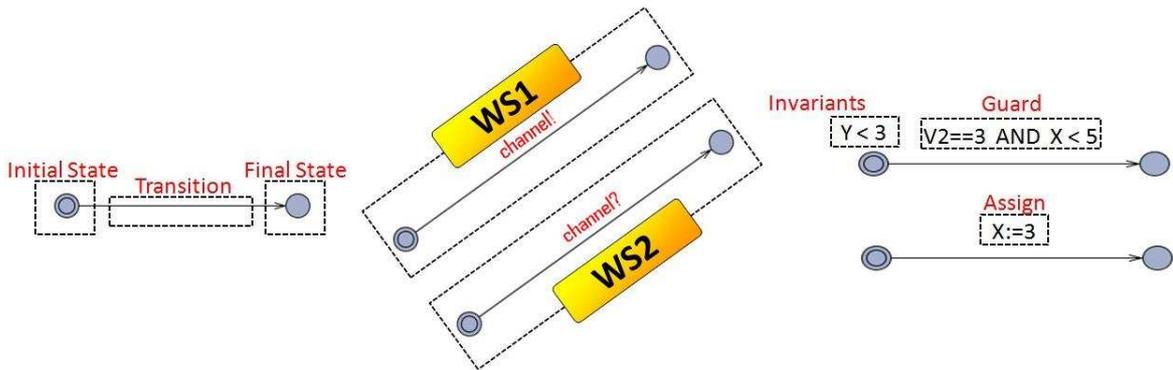


Figure 1. Examples of UPPAAL timed automata.

formalism they use to capture the behaviour of the system and no verification is done in this paper. Other related work is [7] where Gudelj et al. suggest a similar problem to ours. In this approach, they use a Petri net formalism to model the actions to be performed by the actors in this scenario, adding AI techniques (genetic algorithms) as another form of modelling. Indeed, they only show the possible prototype, not making verification. We can mention other related works: [11] and [8]. In [11], WSRF is used to solve the integration problem of various heterogeneous systems in a health information system grid model. In [8], the design and implementation of a Grid-based architecture for scientific workflow is presented. This architecture allows the dynamic discovery of existing Web services in combination to ad-hoc developed ones.

One of the main contributions of this paper is to define a primer version of the necessary elements to model and check Web services with stateful resources. The corresponding translation into timed automata will then be defined. In some previous works, such as [2] and [5], the verification of Web services compositions by means of timed automata has been considered, but without stateful resources. The other main contribution is to develop a scheduling meta-model with one part of the WSRF’s specifications (WS-Notification and WS-ResourceProperties) improving this work with a verification phase, using the UPPAAL model checker.

The rest of the paper is structured as follows: Section II contains the needed background of our approach, that is, the UPPAAL tool, and the WSRF specification. In Section III we specify the elements we need to model and check Web services with stateful resources, and the translation of these elements into timed automata. A case study is included to show how the approach works. Section IV shows how the verification process is carried out over the case study. Finally, Section V contains the conclusions and future work.

II. BACKGROUND

A. UPPAAL

UPPAAL [9] is a tool box for modelling, simulation, validation and verification of real-time systems, based on constraint-solving and on-the-fly techniques, developed jointly by the Uppsala University and the Aalborg University. It is appropriate for systems that can be modelled as a collection of non-deterministic processes with finite control structures and real-valued clocks, communicating through channels and (or) shared variables. Thus, a UPPAAL system consists of a set of concurrent processes, each of which being modelled by a timed automaton. This automaton consists of a set of nodes and a set of transitions. To define the behaviour of the system it is possible to define “invariants”, “guards” and “synchronizations” in the automata:

- The “synchronization” between processes is done through “channels”. One of the processes, which is called the initiator of the synchronization, will invoke the channel with the symbol “!”, while the other process will invoke the channel with the symbol “?”.
- A “guard” is a trigger condition of a transition. It expresses a condition over clocks and integer variables, which must be satisfied when the transition is taken.
- An “invariant” is a condition of progression associated with a node. It indicates the time that the automaton can remain in that node.

Figure 1 depicts some examples of timed automata representations in UPPAAL. On the left-hand side we can see how the states (nodes) of the automata are represented and the transitions between these states. On the centre we can see how two automata representing two different Web services (WS1 and WS2) can be synchronized by means of a channel. Finally, on the right-hand side we can see how invariants, guards and assignments are represented in the automata.

B. Web Services Resource Framework (WSRF)

WSRF [1] is a specification developed by OASIS (Organization for the Advancement of Structured Information

Name	Describes
WS-ResourceLifetime	Mechanisms for WS-Resource Destruction, including message exchanges that allow a requestor to destroy a WS-Resource.
WS-ResourceProperties	Definition of a WS-Resource, and mechanisms for retrieving, changing, and deleting WS-Resource properties.
WS-RenewableReferences	A conventional decoration of a WS-Addressing endpoint reference with policy information needed to retrieve an updated version of an endpoint reference when it becomes invalid.
WS-ServiceGroup	An interface to heterogeneous by-reference collections of Web services.
WS-BaseFaults	A base fault XML type for use when returning faults in a Web services message exchange.

Figure 2. WSRF technical specification.

Standards) and some of the most pioneering computer companies, whose purpose is to define a generic framework for modelling Web services with stateful resources, as well as the relationships between these services in a Grid/Cloud environment. This approach consists of a set of specifications that define a representation of WS-Resource in the terms that specify the messages exchanged and the related XML documents. These specifications allow the programmer to declare and implement the association between a service and one or more resources. It also includes mechanisms to describe the means to check the status of a resource and the service description, which together form the definition of a WS-Resource. Furthermore, they define the necessary steps to make the state of a Web service accessible through its interface (described in WSDL) and related mechanisms to addressing and grouping defined elements in the WS-Resource. WSRF is useful to declare, create, access, monitoring and destroying WS-Resources through conventional mechanisms. These conventional mechanisms are described as follows (Figure 2 summarizes some of them):

- *WS-ResourceLifetime*: The lifetime of a WS-Resource is defined as the period between its instantiation and destruction. The mission of this specification is to standardize the process of destroying a resource and identify mechanisms to monitor this lifetime, but this specification does not define how to create the WS-Resource. It includes two ways to destroy a resource: immediately through an explicit message or timed destruction.
- *WS-ResourceProperties*: WSRF uses a precise specification to define the properties of the WS-Resources. This definition will consist of the definition of the interface in WSDL and an XML document (Resource Properties Document) that specifies the properties of the associated resource, for example, the disk size, processor capacity, etc. If the user wants to access,

modify or update this document it is necessary to use a series of messages defined by the specification.

- *WS-ServiceGroup*: This specification allows the creation of groups that share a common set of properties, i.e., it is a mechanism for grouping together different Web services with similar behaviour.
- *WS-Basefaults*: The developer typically uses a Web service interface defined by others, so a method to standardize the format for reporting error messages facilitates the work. This is the main goal of WS-BaseFaults.
- *WS-Notification*: This specification allows a *NotificationProducer* to send a notification message to a *NotificationConsumer* in two ways:
 - 1) The *NotificationProducer* sends a notification message to the *NotificationConsumer* without following any formalism.
 - 2) The *NotificationProducer* uses a specific formalism to send notifications.

The option selected is sent by the subscriber in the subscription message. Thus, the second option allows the user to receive a wide range of notification messages, but the user can receive many topics in which they are not interested because the information sent in these messages is obtained from a topics tree stored in the Web service.

- *WS-BrokeredNotification*: A *NotificationBroker* is an intermediary, who, among other things, allows interactions between one or more *Publishers* and one or more *NotificationConsumers*. The mission of the *Publisher* is to observe situations and create notification messages to report these situations, while the broker is responsible for forwarding these messages.

III. SERVICE + RESOURCE MODELLING

In this section, we show the necessary elements to model and check Web services with stateful resources and the corresponding translation into timed automata for verification.

Concerning the broker Web service, we need four channels to model the actions that this service can support: *Notification* (to send notifications to the others services), *QueryChannel* (to receive information), *ResponseChannel* (to send information), *PublishChannel* (to publish the information about the topics) and, finally, *SubscribeChannel* (to receive requests of subscription to one or more topics). In addition, we have three variables: *v* that represents the value of the data received or sent, *op* is the operation to perform and *id* is the identifier of the variable. These channels and variables have the same meaning in the Web service automaton. Figure 3 depicts the automaton for broker Web service.

Figure 4 depicts the automaton that models the Web service behaviour. The main difference between this automaton and the previous one is the task assigned to the channels

(receive or send data depending on the direction of the communication).

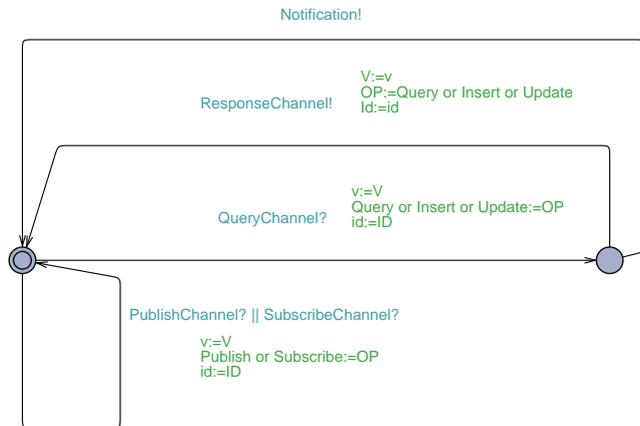


Figure 3. Broker service automaton.

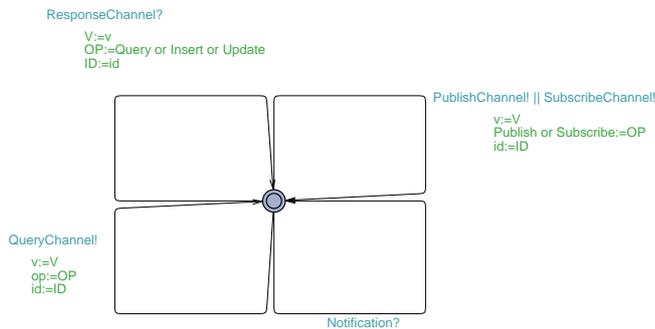


Figure 4. Web service automaton.

A. Case Study: CONTAINER TERMINAL PLANNING

Our modelling problem is a resource allocation for a series of particular tasks. In the case of WSRF, these tasks can be Web services and the resources would be the associated stateful resources. The description of the particular problem is the following: Given a number of trucks (tasks) with designated unloading window (in our case, the time between the WS-Resource creation and its timed destruction), assign the cranes (resources) to them, supposing that our system controls each one of the cranes at the port. When the trucks are near the port (10 Km.), they must report to the crane control tower (the broker role in WSRF) that they want to subscribe to the topic *CraneFree* so they can receive a notification of when they can unload their load. The broker assigns the cranes based on time windows, so it

always chooses the truck with the smaller time window to ensure the system correct behaviour. Once the truck has finished its work, it must send a notification to the broker (*CraneFree*). For simplicity, we will call *Crane* the cranes control tower automaton and *Truck* the truck automaton. Figures 5 and 6 depict the cranes control tower and truck automata respectively.

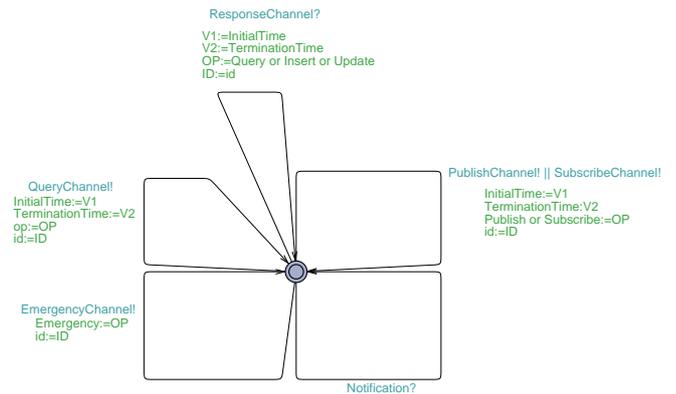


Figure 5. Control tower automaton.

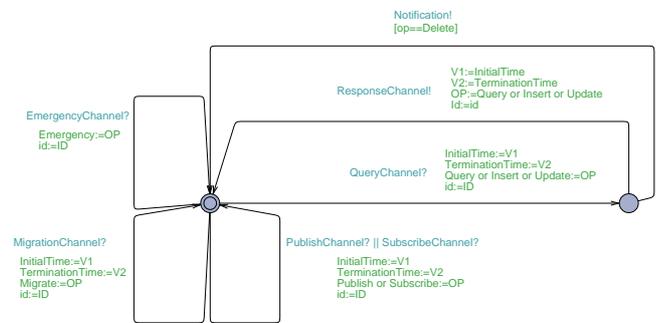


Figure 6. Truck automaton.

Next, we start by explaining the operation of the system in normal conditions and, after that, describing the exceptional behaviour. When a truck approaches to the port, it uses the *SubscribeChannel* to report its arrival time (*InitialTime*), its completion time (*TerminationTime*) and its identifier (*id*). The tower receives the information by its *SubscribeChannel* (so, we use ?) and immediately activates the *QueryChannel* to introduce this information in the database (op == Insert). The crane control tower responds with the information sent to it to find inconsistencies in the stored data. The trucks can query (op == query) and update (op == update) this data at any time by following the same steps.

In the event of receiving an emergency message, the system will insert the truck data at the top of the database (first position to unload) . Then, it sends a notification to the trucks which are approaching to the cranes for checking whether it is possible or not to unload later.

If it is a migration from one crane to another, the control tower will receive requests through *MigrationChannel*, proceeding as in the first case.

Finally, when a truck ends up its work, it notifies this situation to the tower by using *PublishChannel* and it proceeds to remove its information of the database (op==delete). Next, the control tower uses the channel *Notification* to give work permission to another truck.

IV. VERIFICATION

In the previous section, we have shown the translation between the communication model of the container terminal in WSRF and the corresponding timed automata. However, this work would be incomplete without a first approximation of how to model the task scheduling (trucks) in the terminal.

In this section, we present a simple model that represents the system in a general way and after that, we find very important to provide some formulas to check the correctness of the model. In this sense, as noted in the introduction, we use the UPPAAL model checker to ensure deadlock-freeness and search for possible errors to improve our system design. We model the internal behaviour of Web services, i.e., the necessary actions performed by the actors in this scenario to succeed in managing the scarce resources. Note that the figures of this section show a simplified model to ease understandability. The truck timed automaton has been modified to take into account two possible situations: **OnTime** or **Delayed**. Thus, if the unloading of goods is within the time window, the truck is considered on time, while on the other case, it is delayed. On the left-hand side of Figure 7 we show the representation of the *Crane* automaton. This timed automaton uses two channels: request and notification. The first one is used to accept service requests by trucks while the other channel is used to accept the notifications when the trucks end up their work. On the right-hand side, the *Truck* automaton is used to model the different states of the truck.

The next example will help the reader clarify the meaning of transitions and states. Assuming a certain time of arrival t_i , $mint_{early}$ defining the early time in which the truck can arrive to the port and $maxt_{late}$ representing the latter time in which the truck can leave the port, we need to define two possible situations: **OnTime** or **Delayed**. The first one is when the truck can arrive to the port between the interval $[mint_{early}, t_i]$ and the other one is when the truck arrives to the port in $[t_i + 1, maxt_{late}]$. The meaning of the channels is analogous to the *Crane* automaton. In Table I we show the trucks arrival timetable to the port.

TRUCKS	$Mint_{early}$	$Maxt_{early}$	$Mint_{late}$	$Maxt_{late}$
truck1	153	159	160	559
truck2	100	125	126	347
truck3	91	136	137	512
truck4	50	100	101	250
truck5	175	235	236	350
truck6	210	299	300	600

Table I
TRUCKS ARRIVAL TIMETABLE.

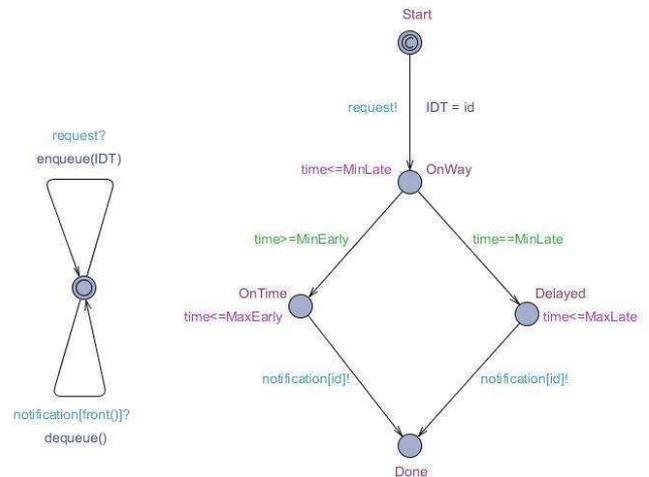


Figure 7. Crane-Truck automata.

To ensure the system correctness, we have formalized the required queries to verify certain properties by using the UPPAAL tool. The first property that we want to check is the absence of deadlocks in the model ($A \square \text{not deadlock}$) and the second is the existence of an execution trace that allows all the truck automata to reach the state *done*, that is, all trucks accomplish their work ($E \langle \langle \text{Truck1.Done and Truck2.Done and Truck3.Done and Truck4.Done and Truck5.Done and Truck6.Done} \rangle \rangle$). In Figure 8, the model checker obtains that the second formula is **satisfied**, so we ask UPPAAL to show us the trace that satisfies the formula, obtaining the trace depicted in Figure 9. Due to space limit, we only show the important part of this trace where we can see the necessary order of notifications to avoid the deadlocks in our model. Based on this we can ensure that the control tower needs to serve the trucks in this sequence: truck4, truck3, truck2, truck1, truck5 and truck6. Besides, the first formula is **not satisfied**, so we can ensure that the model has deadlocks. As we have found a design error in our model, we would have to go back to the design phase, correct the problem and repeat the process to check these properties again. The solution of this error is very simple since we have not taken into account the order of crane requests. The best way to solve the problem is by adding an incoming buffer to store the truck requests and sorting it according to the arrival time.

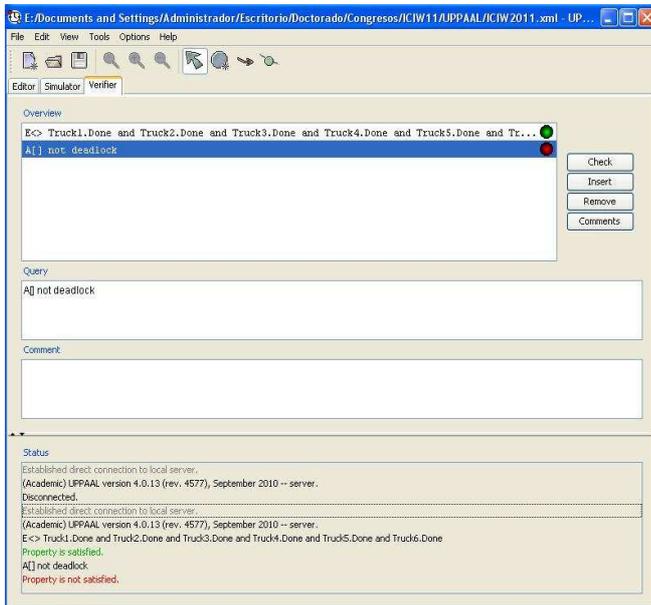


Figure 8. Screenshot of the UPPAAL verifier for Container terminal planning.

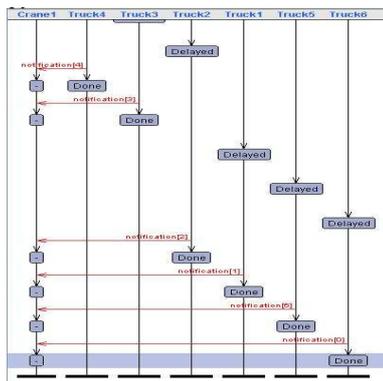


Figure 9. Screenshot of the UPPAAL simulator for Container terminal planning.

V. CONCLUSIONS AND FUTURE WORKS

Using formal methods is always beneficial to model, check and verify computer systems endowing these systems with mathematical rigour, minimum error rate and conformance with the specification. Moreover, adding stateful resources to Web services allows these services to store information that can be used in the future. This paper is a first approximation to the formal verification of WSRF specifications. Thus, we have shown how formal techniques, and in this specific case, timed automata, can be used to model and verify the use of resources in a Web services system. As future work, we are considering the possibility of implementing this model in Globus toolkit 4 by adding directly the queue system needed to ensure deadlock-freeness. Furthermore, we are working on a translation of WSRF into timed automata and Petri nets.

VI. ACKNOWLEDGEMENT

Partially supported by the Spanish government (co-financed by FEDER funds) with the project TIN2009-14312-C02-02 and the JCCLM regional project PEII09-0232-7745.

REFERENCES

- [1] T. Banks. *Web Services Resource Framework (WSRF) - Primer*. OASIS, 2006.
- [2] M. E. Cambroner, G. Díaz, V. Valero, and E. Martínez. *Validation and verification of Web services choreographies by using timed automata*. *Journal of Logic and Algebraic Programming*, vol. 80(1), pp. 25-49, 2011.
- [3] E. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MITPress, Cambridge, 1999.
- [4] K. Czajkowski, D. Ferguson, I. Foster, J. Frey, S. Graham, I. Sedukhin, D. Snelling, S. Tuecke, and W. Vambenepe. *THE WS-RESOURCE FRAMEWORK VERSION 1.0*. <http://www.globus.org/wsrf/specs/ws-wsrf.pdf>, 2004.
- [5] G. Díaz, J. J. Pardo, M. E. Cambroner, V. Valero, and F. Cuartero. *Verification of Web Services with Timed Automata*. *Notes in Theoretical Computer Science*, vol. 157(2), pp. 19-34, 2006.
- [6] I. Foster, J. Frey, S. Graham, S. Tuecke, K. Czajkowski, D. Ferguson, F. Leymann, M. Nally, T. Storey, and S. Weerawaranna. *Modeling Stateful Resources with Web Services*. Globus Alliance, 2004.
- [7] A. Gudelj, M. Krcum, and Dragan Cacic. *Container Terminal Planning by Petri-net and Genetic Algorithms*. *Proceedings of 10th International Conference on Traffic Science (ICTS), Transportation and globalization*, 2006.
- [8] D. Laforenza, R. Lombardo, M. Scarpellini, M. Serrano, F. Silvestri, and P. Faccioli. *Biological Experiments on the Grid: A Novel Workflow Management Platform*. *20th IEEE International Symposium on Computer-Based Medical Systems (CBMS'07)*, pp. 489-494, 2007.
- [9] K.G. Larsen, P. Pettersson, and W. Yi. *UPPAAL in a Nutshell*. *International Journal on Software Tools for Technology Transfer (STTT)*, 1997.
- [10] T. Zang, R. Calinescu, S. Harris, A. Tsui, M. Kwiatkowska, J. Gibbons, J. Davies, P. Maccallum, and C. Caldas. *WSRF-Based Modeling of Clinical Trial Information for Collaborative Cancer Research*. *IEEE International Symposium on Cluster Computing and the Grid*, vol. 0, pp. 73-81, 2008.
- [11] H. Ping and W. Xin-Lei. *Health Information System Grid Based on WSRF*. *Second International Conference on Information Technology and Computer Science*, pp. 518-521, 2010.
- [12] S. Tuecke, K. Czajkowski, I. Foster, J. Frey, S. Graham, C. Kesselman, T. Maquire, T. Sandholm, D. Snelling, and P. Vanderbilt. *Open Grid Services Infrastructure (OGSI) Version 1.0*, 2003.

The Location-based Authentication with The Active Infrastructure

David Jaros, Radek Kuchta, Radimir Vrba

Department of Microelectronics
 FEEC, Brno University of Technology
 Brno, Czech Republic
 jarosd|kuchtar|vrbar@feec.vutbr.cz

Abstract - The paper introduces location-based authentication techniques that are especially addressed to use in buildings and the environment, which is not covered by GPS signal (Global Position System). An active infrastructure is used as a source of position information. Two techniques are proposed. The first technique performs a remote user's authentication where the user's terminal broadcasts identity message. The active infrastructure evaluates user's position and informs an authenticator. The other proposed technique performs local user's authentication. The authentication process is launched on the user's terminal. The user's terminal collects information from its actual neighborhood and evaluates position information.

Keywords - *Location-based authentication, active infrastructures, local authentication, remote authentication*

I. INTRODUCTION

The location-based authentication is a quite new direction in the access management. The direction gains in importance nowadays due to mobile devices coming to wireless network environment. The advantages and a possible application scenario is discussed in [1, 2].

The user's position information is required for access management systems more and more often. The user should bring up information about his/her position with the other credentials when he/she attempts to get access to protected service or content. For example, the user's right in the private company network can be assigned depending on his/her position. The access management system can make decision about the result of user's authentication or can assign set of rights depending on the user's position. The access management system is generally called AAA system (Authentication, Authorization and Accounting) regarding the three main processes covered in [3]. The user's position information could be processed mainly in the authentication and authorization. The authentication techniques that use user's position information are called location-based authentication.

In this paper, we propose two techniques. The first of them is remote authentication. The user accesses to remote network resources in this case and sends his/her credentials over the network. The authentication process proofs brought up credentials on the remote machine (server).

On the other hand, the local authentication is launched on the user's terminal; the user brings up credentials locally. This case is useful when the protected content or services are

stored on user's terminal. The other possible application scenario can be found in authentication during logon to laptop operating system.

We can divide location-based authentication techniques depending on the source of position information into two main groups. The position information can be sourced from the user's terminal (for example GPS enabled) in the first group. The second group covers techniques where the user's position is evaluated by infrastructure (for example GSM network).

We introduce authentication techniques in which the position information is sourced from the infrastructure. The first technique is a remote type and the second one is a local type, as classification above refers.

The rest of this paper is organized as follows. The second section describes technology infrastructure that is used in the next two proposals. The third section introduces a new propose of the location-based authentication that is based on the active infrastructure and solves the remote authentication. The fourth section deals with our second proposal of the location-based authentication techniques. The second proposal is designed especially for local authentication in the user's terminal.

II. ACTIVE INFRASTRUCTURE

The active infrastructure (AI) is a technology background that is used in the two authentication techniques that are described in the next two sections. The key parts of AI are an anchor point, a user's tag and an authenticator. The anchor point is located somewhere from where some of the users want to be authenticated regarding to his/her position. We assume that the position of anchor point is exactly known for the authenticator. On the other hand, the user's tag is assigned to the particular user and it is hard related with his/her identity. User's tag can be a part of user's terminal or an autonomy pocket device. The position of user's tag is proclaimed in terms of proximity between the anchor point and the user's tag. When the user's tag can communicate with the anchor point it means that is nearby.

Figure 1 presents AI's key parts. The anchor point is on known position x_{AP} , y_{AP} , z_{AP} . If the user's tag is in neighborhood it can communicate with anchor point and it means that anchor point's position is similar to the position of user's tag. The similarity between the positions is dependent on the range of transceivers. When the user claims that he/she is on position nearby the anchor point, the

authenticator asks the anchor point if an appropriate user's tag is in the neighborhood. Here should be noted, that for example IQRF [4], Bluetooth [5] or something similar can be used as wireless technologies.

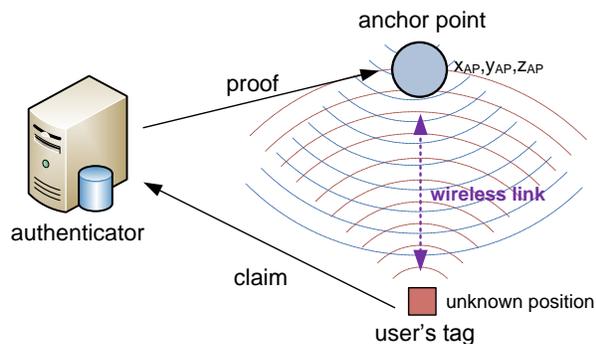


Figure 1. Principle of an active infrastructure

The relationship between the anchor points and the authenticator should be mutually trusted. We propose the use of symmetrical cryptographic system AES (Advanced Encryption System) that is described in [6]. The trusted relationship between communicating parts has to be established before the first use. An initial binding process covers key generation and its exchange. This process has to be granted by system administrator because it is crucial for system security. The binding process between the anchor point and the authenticator is described as follows.

1. First, a secured channel between both sides has to be established. This is provided by the Diffie-Hellmann's principle[7].
2. The authenticator generates AES's key that will be used in the future whenever the authenticator will communicate with this anchor point.
3. The generated key is sent through the secured channel to the anchor point.

The above described AI can vary depending on authentication technique in which it is used.

III. REMOTE AUTHENTICATION

We introduce remote location-based authentication technique with AI in this section.

A possible application scenario for remote location-based authentication is in figure 2. In this scenario user's authentication is processed and evaluated dependent on his/her position. The technique is provided by two independent processes. The first one is user's tag localization; this process is described as follows.

The user's tag is recognized by the anchor point as soon as it is in anchor point's range (A1). The anchor point informs AAA system about this event (A2).The AAA system creates a record in database (A3). Each record contains time, user's identification and anchor point's identification.

The second process is authentication that is initiated by user's requesting of the protected content. The whole process is described below.

1. User's request is redirected to AAA systems that provide access management.
2. AAA system will request credentials from user A.
3. User A will replay with his/her credentials.
4. Part of user's credentials is claimed user's position. The AAA system will query if it is the user who is currently authenticated in claimed position.
5. The AAA system receives answer from the database. When the user is on correct position and his/her position was proved, position condition has been fulfilled.
6. AAA system will inform server with protected content and user's terminal when each of brought up credentials are proved.
7. Access to protected content can be established after authentication and authorization processes are done and when they are correct.

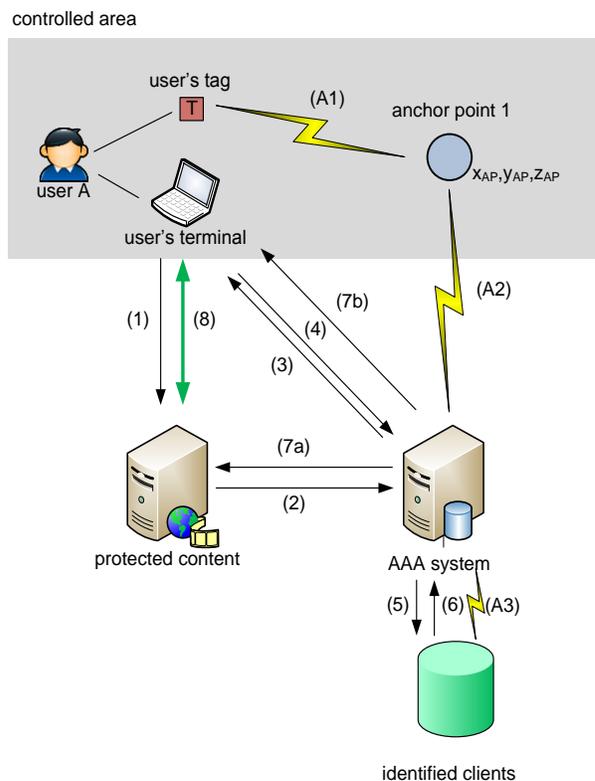


Figure 2. The remote location-based authentication schematic

The remote authentication can be adopted especially to protect sensitive information in private company network. The user has to be in his/her office when he/she wants to work with protected content. The position condition should be periodically tested to prevent user's moving out of

controlled area (out of the office). When a user moves out of controlled area, he/she lose access to protected content. Therefore a time period of re-authentication should be dependent on target application where is technique used.

IV. LOCAL AUTHENTICATION

The location-based technique for local authentication is described in this section. The technique is namely addressed to enhance login process in laptop operating system.

The main difference in comparison to remote authentication is the situation of the authenticator. The authenticator is a part of the user's terminal in this technique. The authenticator has to store a table with anchor point's positions and their encryptions keys, as well as it is in the remote authentication. Initial bonding between anchor points has to be done before the first use, too. In this case the user's tag is a part of authenticator and it can be used by different users. The authentication technique is described in Figure 3. The whole process can be depicted as follows.

When the user tries to log on his/her terminal, the user's tag is activated and it surveys its neighborhood. All available anchor points are captured. The user inputs the identification and other credentials if required. In regard to user's profile, there are processed authentication and assigned right in the authorization.

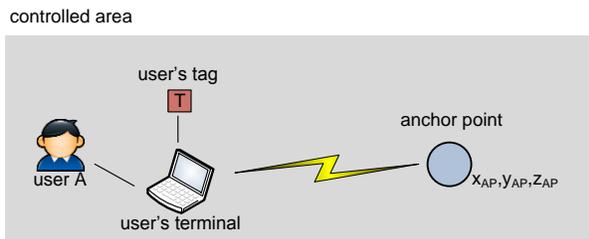


Figure 3. The local location-based authentication

The authentication technique shortly depicted above, could be suitable to assure that data stored on user's terminal will be viewed just in the right place.

V. CONCLUSION

The location-based authentication is a quickly developing field in the access management due to enhancement of mobile devices that are coming into the network environment. We can divide the above mentioned techniques into two basic groups dependent on the source of position information.

The active infrastructure is introduced in the second section. The active infrastructure provides position information of the authenticating user. The key parts of the active infrastructure are the anchor points and user's tags.

The main goal of the article is a proposal of the location-based authentication techniques that are usable in the environment where GPS signal is not available.

In the third section there is an application scenario of the active infrastructure being described. In this scenario the

authentication process runs on a remote machine as it is frequent in the network environment.

The forth section introduced the other proposed technique that is namely addressed to local authentication on user's terminal. In this case an authentication's entity is part of the user's terminal.

The authentication techniques were theoretically proposed till now. The future work will be focused on implementation of the proposed techniques. The active infrastructure test bed should be assembled at first.

ACKNOWLEDGEMENT

This research has been supported by the Czech Ministry of Education, Youth and Sports in the frame of MSM 0021630503 *MIKROSYN New Trends in Microelectronic Systems and Nanotechnologies* Research Project, partly supported by 2C08002 Research Project *KAAPS Research of Universal and Complex Authentication and Authorization for Fixed and Mobile Computer Networks* in the frame of the National Program of Research II, ARTEMIS JU in Project No. 100205 *Process Oriented Electronic Control Units for Electric Vehicles Developed on a multi-system real-time embedded platform*, by ENIAC JU in Project No. 120001 *Nanoelectronics for an Energy Efficient Electrical Car*, partly by the Czech Ministry of Industry and Trade in projects FR-TI1/057 *Automatic stocktaking system* and FR-TI1/058 *Intelligent house-open platform*.

REFERENCES

- [1] D. E. Denning and P. F. MacDoran, "Location-based authentication: Grounding cyberspace for better security," *Computer Fraud & Security*, vol. 1996, pp. 12-16, 1996.
- [2] Karaoguz and Jeyhan, "Location-based authentication of wireless terminal," US Patent, 2011.
- [3] H. Rui, *et al.*, "A novel service-oriented AAA architecture," in *Personal, Indoor and Mobile Radio Communications, 2003. PIMRC 2003. 14th IEEE Proceedings on*, 2003, pp. 2833-2837 vol.3.
- [4] Microrisc. (2011, 30-01). *IQRF homepage*. Available: www.iqrf.org
- [5] B. SIG. (2011, 30-01). *Bluetooth homepage*. Available: www.bluetooth.com
- [6] L. Chi-Feng, *et al.*, "Fast implementation of AES cryptographic algorithms in smart cards," in *Security Technology, 2003. Proceedings. IEEE 37th Annual 2003 International Carnahan Conference on*, 2003, pp. 573-579.
- [7] Y. Eun-Jun and Y. Kee-Young, "An Efficient Diffie-Hellman-MAC Key Exchange Scheme," in *Innovative Computing, Information and Control (ICICIC), 2009 Fourth International Conference on*, 2009, pp. 398-400.

Continuous Evaluation in the process of Ontology Development

Dejan Lavbič, Marjan Kriper, and Marko Bajec

Laboratory for Data Technologies

ULJ, Faculty of Computer & Information Science

Ljubljana, Slovenia

{dejan.lavbic, marjan.krisper, marko.bajec}@fri.uni-lj.si

Abstract—Due to complexity of existing methodologies for ontology development we propose facilitating ontology development with continuous evaluation of steps in the process of ontology development. The approach is called Rapid Ontology Development (ROD) and is based on completeness indicator that helps guiding developer by constant evaluation of ontology and producing recommendations to progress to next step and improve the quality of ontology. The applicability of the approach is demonstrated on Financial Instruments and Trading Strategies (FITS) ontology. The main contribution of the paper is the suggested approach for rapid development of ontologies which brings ontology modeling closer to business users as it does not require from users to know any formal syntax.

Keywords- *Rapid ontology development, Business oriented approach, Ontology evaluation.*

I. INTRODUCTION

The simplicity of using approaches for ontology construction and accompanying tool support is an important issue which needs a lot of attention and further work. Current approaches in ontology development are technically very demanding and require long learning curve and are therefore inappropriate for business users. In majority of existing approaches an additional role of knowledge engineer is required for mediation between actual knowledge that business users possess and ontology engineers who encode knowledge in one of selected formalisms. The use of business rules management approach seems like an appropriate way to simplification of development and use of ontologies in business applications. Besides simplifying the process of ontology creation we also have to focus on very important aspect of ontology completeness. The problem of error free ontologies has been discussed, e.g., in [15] and [7]. One of the goals of Rapid Ontology Development (ROD) approach that authors propose is constant evaluation of ontologies during the development process for major types of errors. User can therefore, based on recommendations, improve the ontology and eliminate the error. It is also a very important aspect that before the usage the ontology itself is error free. Thus we define ROD model that introduces detail steps in ontology manipulation. The starting point was to improve existing approaches in a way of simplifying the process and give user tool support throughout the lifecycle and not to conclude with developed ontology but enable the use of ontology in various scenarios.

The paper is structured as follows: after describing related works in Section II, we continue with the description of the Rapid Ontology Development model (Section 3). The evaluation of the model is provided in Section IV. Finally, in Section V, we give concluding remarks and ideas for future work.

II. RELATED WORKS

Ontology is a vocabulary that is used for describing and presentation of a domain and also the meaning of that vocabulary. The definition of ontology can be highlighted from several aspects, e.g., from taxonomy ([3], [17], [22]) as knowledge with minimal hierarchical structure, vocabulary ([1], [11]) with words and synonyms, topic maps ([6], [14]) with the support of traversing through large amount of data, conceptual model ([10], [12]) that emphasizes more complex knowledge and logic theory ([3], [23]) with very complex and consistent knowledge.

Ontologies are used for various purposes ([2], [4], [5], [9], [16]) such as natural language processing, knowledge management, information extraction, intelligent search engines, digital libraries, business process modeling, etc. While the use of ontologies was primarily in the domain of academia, situation now improves with the advent of several methodologies for ontology manipulation. Existing methodologies for ontology development in general try to define the activities for ontology management, activities for ontology development and support activities. More detailed insight into wide spectrum of methodologies can be found, e.g., in [3], [5], and [19], whilst here only the most representative are depicted. CommonKADS [18] is focused towards knowledge management in information systems with analysis, design and implementation of knowledge. Enterprise Ontology [21] is the groundwork for many other approaches and is also used in several ontology editors. METHONTOLOGY [8] enables the construction of ontologies at the knowledge level and the approach is very close to prototyping. Another approach is TOVE [20] where authors suggest using questionnaires, which is useful where domain experts have very little knowledge of knowledge modeling. OTK Methodology [19] defines steps in ontology development into detail and introduces two processes – Knowledge Meta Process and Knowledge Process. UPON [13] is based on Unified Software Development Process and is supported by UML language. DILIGENT [5] is focused on different approaches to distributed ontology development.

In aforementioned methodologies there is a lack of Rapid Application Development (RAD) approaches in ontology development, the use of ontologies in business applications and approaches analogous agile methodologies in software engineering. There is also an evident lack of approaches that do not require extensive technical knowledge of formal languages and techniques for capturing knowledge from domain experts. The majority of approaches require an additional role of knowledge engineer that transfers the knowledge into formal syntax within knowledge base.

This paper introduces a novel approach in ontology modeling based on good practices and existing approaches while trying to eliminate the need of knowing formal syntax required for codifying the ontology and therefore bringing ontology modeling closer to business users who are actual knowledge holders. The following section will introduce the process, required tasks and highlight the advantages of Rapid Ontology Development (ROD) approach.

III. RAPID ONTOLOGY DEVELOPMENT MODEL

A. ROD process

Ontology development following ROD approach is through 3 stages pre-development, development and post-development as depicted in Figure 1. Every stage delivers a specific output with the common goal of creating functional component based on ontology that can be used in several systems and scenarios. In pre-development stage the output is feasibility study that is used in subsequent stage development to construct essential model definition. The latter artifact represents the schema of problem domain that has to be coupled with instances from the real world. This is conducted in the last stage post-development which produces functional component for usage in various systems.

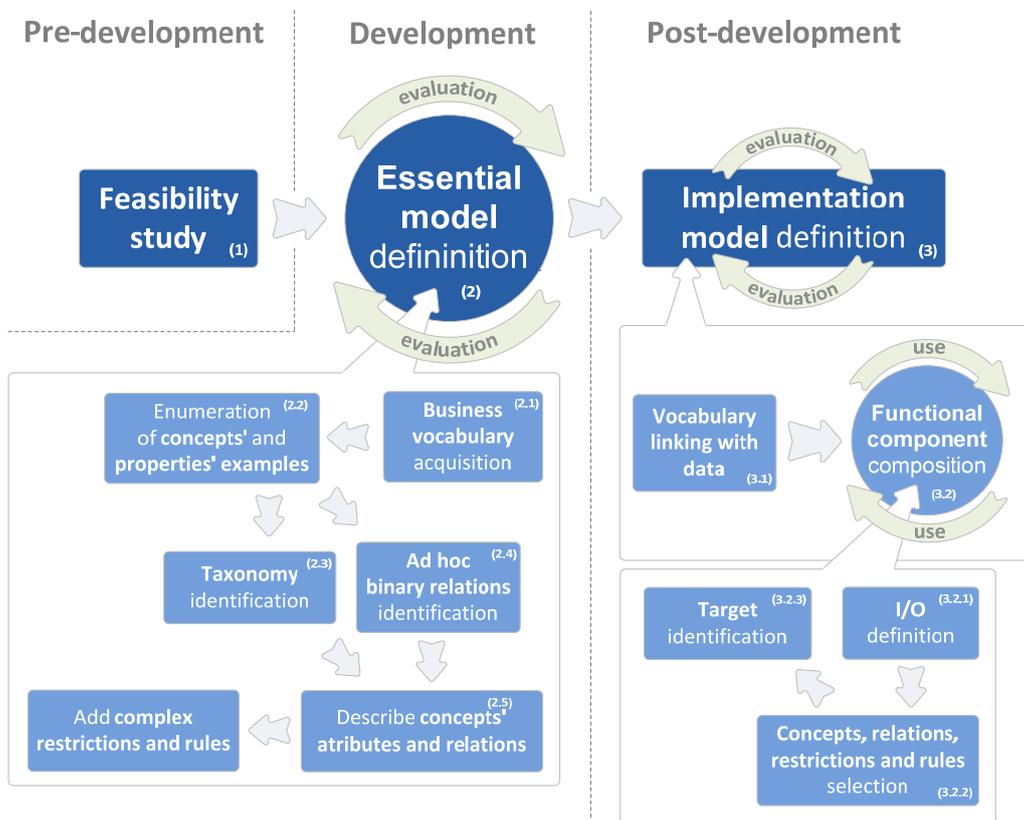


Figure 1: Process of Rapid Ontology Development

The first stage called pre-development is concerned with feasibility study (step 1) of problem domain. This step includes assessing the scope of the project with clear definition of boundaries. Next stage is development with the main goal of producing essential model definition (step 2). The development stage contains several steps: business vocabulary acquisition (step 2.1), enumeration of concepts' and properties' examples (step 2.2), taxonomy identification

(step 2.3), ad hoc binary relations identification (step 2.4), describe concepts' attributes and relations (step 2.5) and add complex restriction and rules (step 2.6). Very important aspect of this stage is constant evaluation of developed ontology using ontology completeness assessment indicator OC which is presented in Section 3.2. The last stage is post-development where implementation model definition (step 3) is constructed. The post-development stage contains 2 steps:

vocabulary linking with data (step 3.1) and functional component composition (step 3.2). The latter step is furthermore decomposed into I/O definition (step 3.2.1), concepts, relations, restrictions and rules selection (step 3.2.2) and target identification (step 3.2.3). From evaluation point of view this stage is similar to development stage, because is also constantly evaluated using ontology completeness assessment indicator OC (see Section III.B).

B. Ontology Completeness

To aid users and to simplify progressing through steps in process of Rapid Ontology Development, ontology completeness indicator is introduced. There are 2 main purposes of OC indicator:

(1.) It can be used independently of ROD process (with other ontology development methodologies or ad hoc). Based on semantic review of ontology, enhancements for ontology improvement are available to the user in a form of multiple actions of improvement sorted by their impact. Besides actions and their impacts, detail explanation of action is also available.

(2.) As a helper tool and facilitator in progressing through steps of ROD process. While the user is in a certain step of the process, the OC measurement is adapted to that step by redefinition of weights for calculation. When OC measurement reaches a threshold (e.g., 80%) user can progress to the following step. The adapted OC value for every phase is calculated on-the-fly and whenever a threshold value is crossed, a recommendation for progressing to next step is generated. This way user is aided in

progressing through steps of ROD process from business vocabulary acquisition to functional component composition. In case that ontology already exists, with OC measure we can place the completeness of ontology in ROD process and start improving ontology in suggested phase of development.

Ontology completeness (OC) is defined as $OC = f(C, P, R, I) \in [0,1]$, where C is set of concepts, P set of properties, R set of rules and I set of instances. OC can further be defined as $OC = \sum_i^n w_i \cdot leafCondition_i$ where n is the number of leaf conditions (see Figure 2) and leafCondition is a leaf condition, where semantic check is executed. For relative weights and leaf condition calculation the following restrictions apply $\sum_i w_i = 1, \forall w_i \in [0,1]$ and $\forall leafCondition_i \in [0,1]$. Relative weight w_i denotes global importance of leafCondition_i and is dependent on all weights from leaf to root concept.

The tree of conditions in OC calculation is depicted in Figure 2 and contains semantic checks that are executed against the ontology. The top level is divided into TBox, RBox and ABox components. Subsequent levels are then furthermore divided based on ontology error classification [7]. Aforementioned sublevels are description, partition, redundancy, consistency and anomaly. This proposed structure can be easily adapted and altered for custom use. Leafs in the tree of OC calculation conditions are implemented as semantic checks while all preceding elements are aggregation with appropriate weights as depicted in Figure 2.

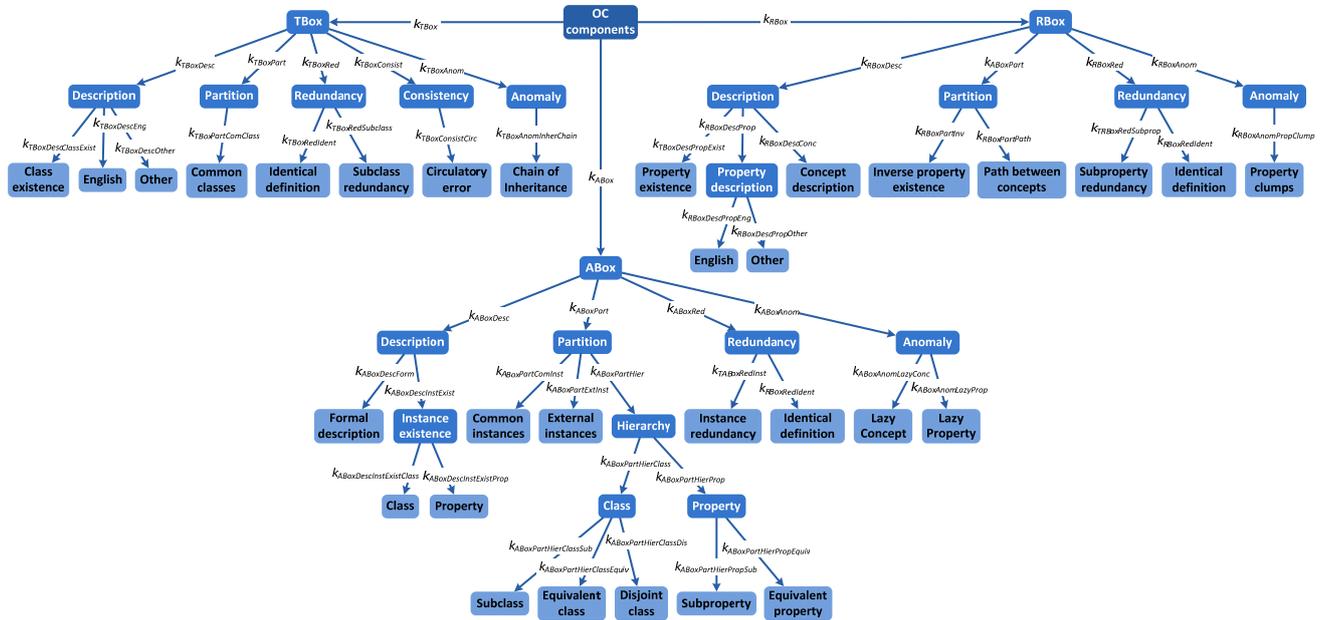


Figure 2: Ontology completeness (OC) tree of conditions, semantic checks and corresponding weights

Certain phases put emphasis on selected components. In initial stages user has to deal with description and structure of concepts, while at the end of essential model definition

restrictions and rules become more important. As post-development and implementation model definition is concerned users have to think about instances of schematic

part of ontologies, therefore attention is shifted to ABox component. There are two types of outputs from ontology completeness (OC) calculation (see Figure 3): (1) OC price and (2) Recommendations to improve OC price.

As depicted in Figure 3 OC price is presented as a value expressed in percentage (e.g., 68%) and visualized as a progress bar. Besides this basic view it turns out that radar chart view of OC price is also very informative as it highlights which areas need improvement. In radar chart view top levels for visualization become description, partition, redundancy, consistency and anomaly as they are more suitable for business users than TBox, RBox and ABox components. All the recommendations are listed in a table view and sorted by their impact. When a recommendation is selected the impact is also depicted in radar chart for better understanding of how the change will affect ontology.

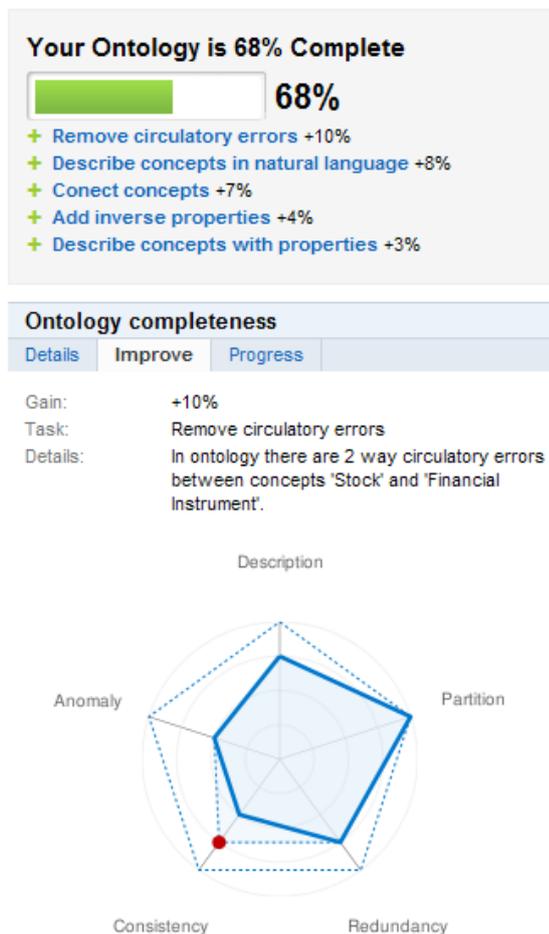


Figure 3: User interface and display of ontology completeness results and improvement recommendations

IV. EVALUATION

Rapid Ontology Development (ROD) process was verified on a case study from financial domain. The requirement was to develop an ontology that contains basic

information about financial instruments, trading and custom trading strategies using ROD. The solution enabled the user to test multiple trading strategies on real trading data that was available for selected stocks in pre-prepared CSV files and Yahoo! Finance resource for real-time and historic data was also available. The developed solution was exported as a functional component (standalone J2SE application) with input parameter of financial instrument symbol and as output trading days that have either buy or sell signals and trade reason. Business user defined the ontology containing required knowledge about financial instruments and at least 3 trading strategies that can easily be combined and tested individually or together. During the process of ontology building user actions were continuously monitored. The grain of collected data was one iteration step where ontology completeness prices and number of ontology elements (concepts, properties and axiom, including rules) were recorded.

The ontology produced is depicted in Figure 4 and is based on 2 simple facts about trading with financial instruments: (1) financial instrument is traded at a stock exchange market and (2) financial instrument is traded on a trading day. Specialization of financial instrument was introduced with Stock, ETF (Exchange Traded Fund) and K.O. certificate. The same approach was used for representing distinction between trading days with buy and sell signals. All concepts were according to OC rules described with formal properties and also natural language.

Trading strategy was implemented in separate ontology that used dynamic import of aforementioned ontology. This was utilized due to modular approach and the ability to develop and use several strategies separately or combined. Trading strategy mainly contains restrictions and rules, while in more complex definition (Japanese candlestick strategy) additional concepts were also introduced. During the experiment 3 different trading strategies were defined: simple trading strategy, simple moving average strategy and Japanese candlestick strategy. Input trading data consisted of quotes data from Apple Inc. (AAPL), Google Inc. (GOOG) and UltraShort S&P500 ProShares (SDS) in a 1 month period. Instances were imported at runtime from World Wide Web (Yahoo! Finance) and File system text files (CSV).

The process of ontology creation and exporting it as functional component is depicted in Figure 5. Chart represents ontology completeness price and number of ontology elements regarding to iterations in the process. During the process of ontology construction based on ROD approach the user was continuously supported by ontology evaluation and recommendations for progressing to next steps. When user entered a phase and started performing tasks associated with the phase (detail description is given in Section III.A), ontology completeness was evaluated as depicted in Figure 1 and further presented in Section III.B. While OC was less than 100% user followed instructions for improving ontology as depicted in Figure 3. Results of OC evaluation are available in a simple view, where basic statistics about ontology is displayed (number of concepts, properties, rules, individuals, etc.), progress bar depicting completeness, and details about evaluation, improvement

recommendations and history of changes. The core element is progress bar that denotes how complete ontology is and is accompanied with a percentage value. Following are recommendation for ontology improvement and their gains (e.g., remove circulatory errors (+10%), describe concepts in natural language (+8%), connect concepts (-+7%), etc.).

When improvement is selected (e.g., remove circulatory errors) the details are displayed (gain, task and details). As depicted in Figure 3 circulatory error can be eliminated with removing the 2 way connection between concepts ‘Stock’ and ‘Financial instrument’ and by doing that gaining 10% in ontology completeness.

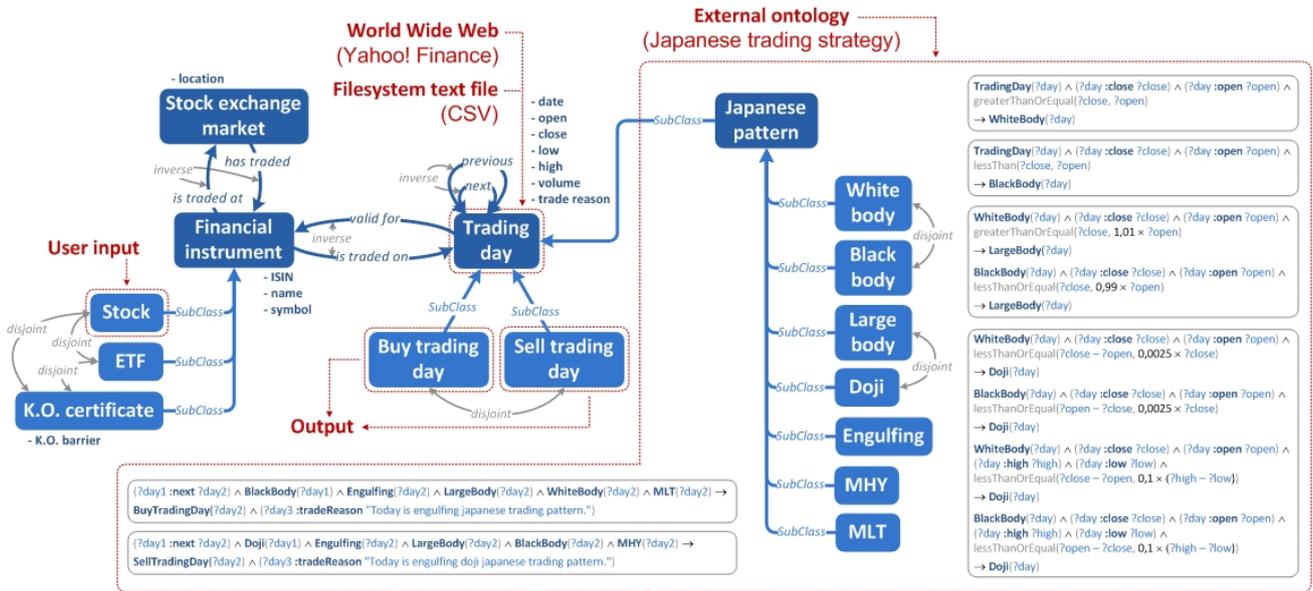


Figure 4: Trading ontology example



Figure 5: Ontology completeness assessment and number of ontology elements through phases of ROD process

The improvement and planned actions are also clearly graphically depicted on radar chart (see Figure 3). The shaded area with strong border lines presents current situation, while red dot shows TO-BE situation if we follow selected improvement. When OC price crosses a threshold value (in this experiment 80%, as depicted in Figure 5) a recommendation to progress to a new phase is generated. We can see from our example that for instance recommendation to progress from phase 2.5 to phase 2.6 was generated in 20th iteration with OC value of 91,3%, while in 19th iteration OC value was 76,5%. As Figure 5 depicts ontology completeness price and number of ontology elements are displayed. While progressing through steps and phases it's

seen that number of ontology elements constantly grows. On the other hand OC price fluctuates – it's increasing till we reach the threshold to progress to next phase and decreases when entering new phase. Based on recommendations from the system, user improves the ontology and OC price increases again.

V. CONCLUSION AND FUTURE WORKS

Available methodologies and approaches usually require very knowledgeable users and developers, while authors propose ROD approach that is more suitable for less technically knowledgeable users. With ROD approach and accompanying IntelliOnto tool business users get an

instrument for ontology modeling that doesn't require very extensive knowledge of ontology languages but still follow and utilize the possibilities of Semantic Web vision. It has been demonstrated on a case study from financial trading domain that a user can build Semantic Web application for financial trading based on ontologies that consumes data from various sources and enable interoperability. The solution can easily be packed into a functional component and used in various systems. By following ROD approach for building Semantic Web applications against existing approaches following advantages can be highlighted: (1) the required technical knowledge for ontology modeling is decreased, (2) the process of ontology modeling doesn't end with the last successful iteration, but continues with post-development activities of using ontology as a functional component in several scenarios and (3) continuous evaluation of developing ontology and recommendations for improvement.

REFERENCES

- [1] Bechhofer, S. & GOBLE, C., 2001, Thesaurus construction through knowledge representation. *Data & Knowledge Engineering*, 37, pp. 25-45.
- [2] Brambilla, M., Celino, I., Ceri, S. and Cerizza, D., 2006, A Software Engineering Approach to Design and Development of Semantic Web Service Applications. In 5th International Semantic Web Conference, Athens, USA.
- [3] Corcho, O., Fernandez-Lopez, M. and Gomez-Perez, A., 2003, Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data & Knowledge Engineering*, 46, pp. 41-64.
- [4] Dahlgen, K., 1995, A linguistic ontology. *International Journal of Human-Computer Studies*, 43, pp. 809-818.
- [5] Davies, J., Studer, R. and Warren, P., 2006, *Semantic Web Technologies - trends and research in ontology-based systems* (Chichester, England: John Wiley & Sons).
- [6] Dong, Y. and Li, M. S., 2004, HyO-XTM: a set of hyper-graph operations on XML Topic Map toward knowledge management. *Future Generation Computer Systems*, 20, pp. 81-100.
- [7] Fahad, M. and Quadir, M. A., 2008, Ontological errors - Inconsistency, Incompleteness and Redundancy. In International Conference on Enterprise Information Systems (ICEIS) 2008, Barcelona, Spain.
- [8] Fernandez-Lopez, M., Gomez-Perez, A., Sierra, J. P. and Sierra, A. P., 1999, Building a chemical ontology using Methontology and the Ontology Design Environment. *Intelligent Systems*, 14.
- [9] Heflin, J. and Hendler, J., 2000, Searching the web with SHOE Artificial Intelligence for Web Search, pp. 36-40 (Menlo Park, USA, AAAI Press).
- [10] Jovanović, J. & Gašević, D., 2005, Achieving knowledge interoperability: An XML/XSLT approach. *Expert Systems with Applications*, 29, pp. 535-553.
- [11] Miller, G. A., 1995, WordNet: a lexical database for English. *Communications of the ACM*, 38, pp. 39-41.
- [12] Mylopoulos, J., 1998, Information modeling in the time of the revolution. *Information Systems*, 23, pp. 127-155.
- [13] Nicola, A. D., Navigli, R. & Missikoff, M., 2005, Building an eProcurement ontology with UPON methodology. In 15th e-Challenges Conference, Ljubljana, Slovenia.
- [14] Park, J. and Hunting, S., 2002, XML Topic Maps: Creating and Using Topic Maps for the Web (Boston, USA: Addison-Wesley).
- [15] Porzel, R. and Malaka, R., 2004, A Task-based Approach for Ontology Evaluation. In ECAI 2004, Workshop on Ontology Learning and Population, Valencia, Spain.
- [16] Rao, J., Dimitrov, D., Hofmann, P. nad Sadeh, N., 2006, A Mixed Initiative Semantic Web Framework for Process Composition. In 5th International Semantic Web Conference, Athens, USA.
- [17] Sanjuan, E. and Ibekwe-Sanjuan, F., 2006, Text mining without document context. *Information Processing & Management*, 42, pp. 1532-1552.
- [18] Schreiber, G., Akkermans, H., Anjewierden, A. et al., 1999, *Knowledge Engineering and Management - The CommonKADS Methodology* (London, England: The MIT Press: Cambridge, Massachusetts).
- [19] Sure, Y., 2003, *Methodology, Tools & Case Studies for Ontology based Knowledge Management* Institute AIFB, pp. 332 (University of Karlsruhe).
- [20] Uschold, M. & Grueninger, M., 1996, *Ontologies: principles, methods and applications*. *Knowledge Sharing and Review*, 11.
- [21] Uschold, M. & King, M., 1995, Towards a methodology for building ontologies. In Workshop on Basic Ontological Issues in Knowledge Sharing (IJCAI '95), Montreal, Canada.
- [22] Veale, T., 2006, An analogy-oriented type hierarchy for linguistic creativity. *Knowledge-Based Systems*, 19, pp. 471-479.
- [23] Waterson, A. and Preece, A., 1999, Verifying ontological commitment in knowledge-based systems. *Knowledge-Based Systems*, 12, pp. 45-54.

A Community Cloud: Archive Retrieval in Multiple Language Services

Wei-hua Zhao¹ Zhan-wei Liu² Zheng-xu Zhao³

Shijiazhuang Tiedao University
Hebei, Shijiazhuang, P R China

1. e-mail: zhaowei9@hotmail.com

2. e-mail: hoson@126.com

3. e-mail: zhaozhengxu@staff.stdu.edu.cn

Abstract—This paper presents a common method and practice in implementing a community cloud, where archives have to support the services in multiple languages and cultures. It demonstrates an internationalization process, the development of software, the migration of archive data and the integration of archive retrieval system. The paper shows how such archive retrieval is carried out and how the system is implemented. Although it does not claim to be the best or the only possible solution, the method provides practical and useful tools for establishing and managing digital archives that serve as a community cloud where it essentially requires internationalization or globalization or localization. The main achievement of the paper resides in the presentation of a complete and useful method for archive retrieval in multiple languages.

Keywords—community cloud; digital archive; globalization; internationalization; information organization

I. INTRODUCTION

Cloud computing is considered as web-based processing, whereby shared resources, software, and information are provided to computers and other devices such as smart phones on demand over the Internet [1]. A community cloud may be established where several organizations have similar requirements and seek to share infrastructure so as to realize some of the benefits of cloud computing. Examples of community cloud include Google's "Gov Cloud" [2]. It could be interesting to consider that archive services are evolving from their traditional form toward an emerging digital domain as a community cloud. Literatures [3,4] reveal that a phenomenal proliferation of digital data clearly underscores the ease with which it is being produced and the capacity in which it is to be accessed. While effective archiving and retrieval the information still remain a work in progress, the importance of and the challenge for accessing it in multiple languages with multiple cultures are being recognized and thus there have been a various efforts made to address the needs in specific areas like education services, government administration and enterprise management.

Archive services across different communities and cultures have evolved into a massive digital media and files in various forms and different languages, yet the current data retrieval practices, due to the limitation of software tools and archive systems, still confine their purposes within specific languages and culture orientations. Those practices in nature

are normally carried out in the native or original forms of the archived documents. For example, the archive system deployed in the Shijiazhuang Tiedao University holds the live data records, predominately in Chinese, of staff and students for over 60 years span in hundreds and thousands of individuals. However each year there have been an increasing amount of service demands for retrieval, translation and interpretation of those native documents into various languages to suit for offshore business, overseas collaborations and of course for those whose are to work and study abroad. The workload in transforming those documents into its specific needs and purposes poses a phenomenal challenge both in time and accuracy [5]. This paper reports an investigation into the method that retrieves, translates and interprets the native documentation into multiple languages and cultures to cater for different service needs. The objective is to establish a useful practice that can utilize various tools currently available to transform the archives from their native form into different languages and culture needs. It is expected that this practice will be an essence for effective data translation, index and migration, thereby providing technical solutions to the aforementioned problems.

The text to follow gives a brief description of internationalization in Section II. Section III presents the use of methodology management in the system design and Section IV details the functions of the system. Section V provides the implementation of the system which is followed by conclusions and future works included in Section VI.

II. INTERNATIONALIZATION

Globalization, internationalization and localization are terms referring to the process of developing software systems, media products, documentation and other collateral material for people who speak foreign languages or constitute a specific cultural group with a large language community. They are long words and people took the habit of writing abbreviations instead. As a result, the three terms are abbreviated as g11n, i18n, l10n, with the intervening number referring to the number of letters between the first and last of each word [6].

A. Terminology

By i18n, it refers to the process by which a program or a set of programs turned into a package is made aware of and

able to support multiple languages. This is a generalization process, by which the programs are untied from calling only English strings or other English specific habits, and connected to generic ways of doing the same, instead. Program developers may use various techniques to internationalize their programs. Some of these have been standardized. GNU Gettext offers one of these standards [7] and this paper will be based on GNU Gettext.

By i18n, it means the operation by which, in a set of programs already internationalized, a program is given all needed information so that it can adapt itself to handle its input and output in a fashion which is correct for some native language and cultural habits. The formal description of specific set of cultural habits for some country, together with all associated translations targeted to the same native language, is called the *locale* for this language or country. Users achieve localization of programs by setting proper values to special environment variables, prior to executing those programs, identifying which locale should be used [7].

For globalization, or g11n, one could consider that it stands for a much broader activities than for some specific technology areas such as software design and system development.

B. Development of i18n

Organizational bodies and commercial companies are always looking for competitive strengths and improved business practices. As Perrow [8] stated, in today's economy, this translates to a demand for better software. The increased

demand has to do with a growing recognition that, throughout the new economy, software is the means for conducting business, tapping new markets globally, and connecting suppliers, manufacturers and end users in a worldwide domain. While the Internet serves as a vital tool for transaction and communication, the software systems themselves - often on either end of an Internet connection - do the heavy lifting and represent the greatest challenges and opportunities for business growth.

An i18n process represents the way in developing such modern software systems that enable organizations to exploit strategically localized advantages and economies of scale to leverage a competitive edge world-wide. The concept of i18n describes the establishment of a network of cross-border activities between and within companies in which any or all of the organizations' departments may be involved. The main standard motives for i18n are seeking market, resource, efficiency and strategic asset [9,10,11]. By any measure, i18n is not a solved problem. Although i18n has become a mainstream software development process, achieving it in streamlined, flexible and standardized manner remains a grand challenge [12]. This may be partially due to the misinterpretation of certain problem areas of i18n [13]. Table I lists the key problem areas and the related questions that are extracted from the discussion, with each area being represented by a question to highlight the fundamental problems for the i18n practice.

TABLE I. PROBLEM AREAS PERTAINING TO THE SOFTWARE DEVELOPMENT PROCESS OF I18N

Problem Area	Description
User interface	Does i18n mean that software can be simply translated via externalising its user interface?
Software translation	Can software translation in i18n process always use the best phrase in the target language?
Programming tool	Can programming tools that themselves have i18n always produce i18n code?
Unicode support	Is software that supports Unicode an i18n system or product by itself?
Open source	Does the use of open source in software product mean that i18n requirements are not applicable for the product?
Standard encoding	Is a standard encoding is automatically making a product having i18n?
Internal tool	Is it true that all company employees speak English so only English needs to be supported by internal tools?
Administration interface	Do administration interfaces need i18n?
Product module or component	Does every product, module or component need i18n?
Product version	Does it mean by adding i18n in the last release a job well done?
Customer feedback	If something is wrong, will the customers' feedback to the developers?
Multilingual capability	Can a software product that works in a foreign language be considered to have i18n?
Base code	Is i18n implemented after the base product is written by a separate group of engineers?
Software engineering	Is i18n only needed in the software development department?

III. METHODOLOGY MANAGEMENT

Methodology management, normally called design methodology management (DMM), originated from computer aided design frameworks that are software environment that integrate design tools and programs required for prototyping computer operating systems and managing the data being generated [14]. To attain rapid design processes of high level automation, frameworks have to select and execute tools automatically for lower-level

tasks to enable designers to concentrate on higher-level decisions. The sequence of design tools is called methodology and the functionality of selecting and executing the tools is called methodology management. A software environment, often part of a framework for selecting and executing design methodology, is called methodology management system [15].

As the design automation community begins to understand the benefits of the technology, its expectations grow: less error-prone design, rapid prototyping systems,

new and customer-tailored product development, high product quality and improved productivity. A good DMM system can bring these benefits to different categories of users such as product designers, tool developers, system developers, chief designers and company managers [14]. DMM is relevant not only to electrical design, but also to software design and other fields like mechanical design. Nevertheless it has not yet become a subject of much discussion in either research community [16] or i18n process. This article introduces DMM into i18n practice to develop an i18n framework and it does so for three purposes. The first one is to demonstrate how an integrated and streamlined i18n process can be carried out and thus to give an insight into the i18n process from an implementation perspective and with sufficient technical details. The second is to provide a practical approach with useful techniques and tools for packaging of software with complete i18n support. The last one is to help in understanding the i18n process in relation to the questions in Table I.

IV. INTERNATIONALIZATION OF ARCHIVE RETRIEVAL

Operations for i18n are often accomplished using software tools, both interactive and automatic, and DMM addresses the need to manage the manner in which these tools are executed to achieve a desired function. According to Fiduk et al [17], this paper adapts the definition of following terms.

1) *Execution environment*: it is a computing environment to manage the tools, tasks, flows, data and information movement that are essential for an i18n process to be accomplished.

2) *Tool*: this is a single executable program capable of performing a specific function toward i18n.

3) *Task*: it is an abstraction of a function toward i18n, for example, translation of an error message or version check of a source code file.

4) *Flow*: it is the order in which tasks are executed. The definition and manipulation of flows provide a mechanism to describe a sequence of tools that make up a process or task and tasks make up a methodology.

5) *Process*: this is a specific combination of tools and/or other processes that performs a function toward i18n.

6) *Tool invocation*: This is the selection of a tool and the use of it to perform what is needed to be done.

7) *Operation*: it is an atomic action within a process.

8) *Methodology*: this is a specified sequence of tasks.

A. The Framework

In general, DMM should have an execution environment that is responsible for user interaction, launching tasks, monitoring processes, automatically executing flows, and so on. For this paper, the execution environment has following specifications:

- It is operated under the common MS Windows XP operating system.
- All tools and software are run as Win32 programs (similar tools and software with different

compilations should be used for other operating systems).

- The i18n development language is C/C++ under Microsoft Visual Studio .NET 2003.
- The i18n software to be developed is open source, so are all internal tools for the i18n process.
- The execution environment is compliance to open source community’s Native Language Support Library and Tools, GNU Gettext Tools, Version 0.14.4 [7].

Based on the above specification, the execution environment is set up and its flow graph is shown in Fig.1. The graph formalism represents individual methodologies. It is actually a bipartite cyclic directed graph that has two types of nodes wherein all edges connect one type to the other and there are no paths from a node back to itself. The two types of nodes are task nodes and specification nodes. Each task node is labelled with a task description. There are two types of task nodes: terminal and non-terminal. A terminal task node represents a run of an i18n tool or programme and is called a tool invocation; it is drawn with single circle.

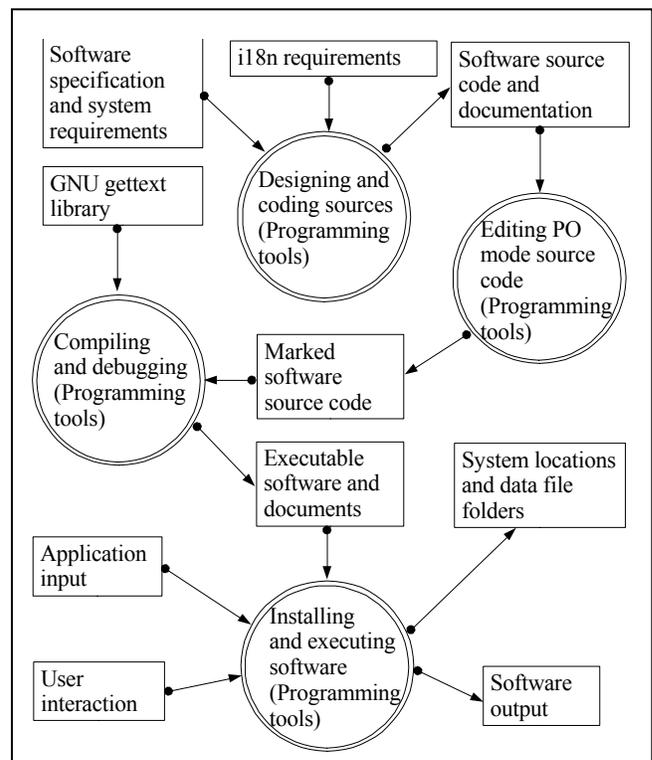


Figure 1. DMM information flow graph of the execution environment for i18n of swar.

The graph representation of the execution environment is arranged in a top-down manner by applying transformations to nodes that represent non-terminal tasks. Fig. 2 shows the next level graph (or sub-graph) that describes a process where the *Marked software source code* is inputted to the *Extracting translatable code* task by which a portable object

template (POT) file named as *PACKAGE.pot* is generated through the *xgettext* tool. The *Comparing and refreshing* task is to obtain updated *Target language PO files*. Note that a target language PO file is a PO file that has the translation of a target language for all original program strings (in file *PACKAGE.pot*); it is generally named as *LANG.po* where *LANG* will be replaced by an ISO639-1 language code [18] when it must refer to the name a specific target language PO file. For example, *ru.po* is the name of a Russian language PO file.

Then the *Comparing and refreshing* task executes the *msgmerge* tool to refresh an already existing *Target language PO file* by comparing it with the up-to-date *PACKAGE.pot* file. The *Source code translating* task is similar to the *Editing PO mode source code* task in the flow diagram shown in Fig. 1, but it is a task of translating the PO files into the target language PO files using *poEdit* tool (see more description about the *poEdit* tool below in Section C). The *Generating MO files* task turns the target language PO files into files of machine-oriented (MO) format.

Finally, the marked software sources code is compiled and linked with the GNU Gettext libraries. This task is automated with Perl [19] scripts managed by the *Perl* tools. This will result in an executable software installed somewhere that users will find it.

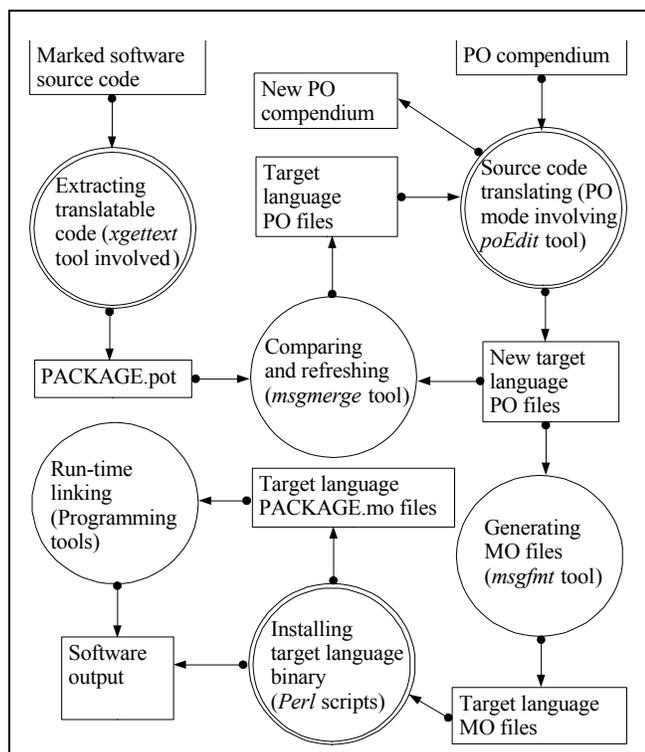


Figure 2. DMM information flow sub-graph of the execution environment for i18n of software.

B. The Archive Retrieval Tools

The archive retrieval a tool is the building block of the framework. With the tools, the framework must define how

to use them and in what order to use them. The first refers to task and the second to flow. The tools in each category are described as follows.

- **Programming Tools:** These are found with Microsoft Visual Studio .NET 2003, mainly the C/C++ compiler and linker. Programming tools includes creating and compiling source code and debugging and converting the source code into libraries, components and executables.
- **Internal Tools:** These are mainly the GNU Gettext utilities. Once the GNU Gettext is installed in the execution environment, these tools are available for the i18n framework to use. The i18n framework reported in this paper only uses following internal tools and invoke each tool with standard commands.
 - a) *xgettext* tool creates the PO template file *PACKAGE.pot*.
 - b) *msginit* tool creates a new target language PO file *LANG.po* from the PO template file *PACKAGE.pot*.
 - c) *msgmerge* tool updates an existing target language PO file *LANG.po* based on a newer version of the PO template file *PACKAGE.pot*.
 - d) *msgfmt* tool generates binary MO files.
- **Manipulation Tools:** GNU Gettext also provides a range of manipulation tools to manipulate PO files in a way that is better performed automatically than by hand. The i18n framework does not use these PO file manipulation tools. Instead, it uses following two open source tools: (1) *poEdit* is a cross platform message catalogues editor which is software for manipulation and translation of PO files. (2) *SciTE* is a text editor that is specialized in source code manipulation and it is used in the i18n framework for text file editing and for manipulating source code and Perl scripts. (3) Another set of tools used is from Perl [19] which creates and runs Perl scripts to automate the i18n tasks such as *Generating MO files* and *Installing target language binary*.

C. Archive Translation and Interpretation

The definition and manipulation of tasks enable a framework to carry out planning without users knowing the details of operations and how they are implemented. Tasks are performed by invoking specific processes. A task could be, for example, encoding Unicode to transform character sets for supporting multilingual languages. For the i18n framework as illustrated in Fig.1 and Fig. 2, the tasks to be performed are non-terminal and terminal. The non-terminal tasks include:

- 1) *Designing and coding sources* is a combination of tasks and processes for creating and programming source code. Tools used in this task are programming tools.
- 2) *Editing PO mode source code* marks the source code according to the GNU Gettext convention toward i18n.
- 3) *Compiling and debugging* is normally to generate the source code into executable software and components which do not necessarily have i18n.

4) *Installing and executing software* is to arrange the software, data files, linked libraries, the source files and documentation in organized filing structures so that they can be accessed by all i18n tools.

5) *Extracting translatable code* is to find translatable code or strings from all source files and then to generate the PO template file. This task mainly uses the *xgettext* tool.

6) *Source code translating* performs all translation of the marked code and strings in the source code from the common language (English) to the target languages. It is carried out using *poEdit* tool.

7) *Installing target language binary* is automated with Perl scripts. This task compiles each target language source code and then links with a MO file into a linked library (dynamic linked library) and installs the results in a specific filing location.

The i18n framework has following three terminal tasks:

1) *Comparing and refreshing* uses the tool *msgmerge* to update PO files whenever the source code is changed.

2) *Generating MO files* uses *msgfmt* tool to generate MO files from target language PO files. This task can be integrated with the *Installing target language binary* task using Perl scripting to form an automated process.

3) *Run-time linking* is carried out during testing and debugging phase of the final software. It will execute the software by run-time linking the library generated by the *Installing target language binary* task for a specific target language. For CJK (Chinese, Japanese and Korean) and other Unicode languages, appropriate font files must be made available for this task to accomplish i18n (generating font file is beyond scope of this paper).

D. The Retrieval Process

There are three general processes involved in the i18n framework. The first one is for preparing source code; the second is for translating the source code; the third is for run-time generating linked library. Each process is of course a combination of tasks and sub-processes.

The first process includes creating and marking the source code. This process is formed by the tasks in Fig. 1. The second process involves in creating and updating the template PO file and editing, updating and manipulating PO files. The third process generates for each target language PO file a target language MO file and then creates a run-time link library for that target language. After the source code is compiled into executable software, this run-time link library enables the software with i18n. This process is the combination of tasks *Installing target language binary*, *Generating MO files* and *Run-time linking* as shown in Fig. 2.

V. SYSTEM IMPLEMENTATION

The framework shows how the complete i18n processes could be carried out. It is certainly not the only possible solution, but it provides a skeleton for real i18n packaging, i.e. the key tasks that have to be accomplished.

A. System Requirements

The key requirement is a version of GNU Gettext Tools (Version 0.14.4 for this paper) which is essential for i18n.

All other internal tools (see the above Section IV) must be installed as part of the execution environment. It must be noted that *Perl* tools should be installed and program compiler and linker tools should also be made accessible in command line mode if *Perl* tools and Perl scripts are used to automate some of the i18n tasks in the i18n framework.

B. Source Preparation

This is about programming or creation of software source code. In this paper, it is C/C++ source. Bringing GNU Gettext convention into software package is to identify in the sources those strings which are meant to be translatable and those which are untranslatable. Beside this, some simple and standard changes are needed to initialize the GNU Gettext library. Changes to the source code fall into three categories. First, the programmer has to make the localization functions known to all modules needing message translation. Second, the programmer should properly trigger the operation of GNU Gettext library when the program initializes, usually from the main function. Last, the programmer should identify and especially mark all translatable strings in the source code [7].

As illustrated in Fig. 1, from the *Designing and coding sources* task to the *Marked software source code* specification, all work should be achieved by programming tools. From this point, the source code is marked according to the GNU Gettext programming convention and can be either compiled into package without i18n or brought to the next stage for translation toward i18n (see Fig. 2).

C. Generating PO template File

Once the source code has been modified and marked, the *Extracting translatable code* task makes invocation of the *xgettext* tool. This tool will find and extract all translatable strings from the source code files and then create the PO template file *PACKAGE.pot*. Below is a typical example of the invocation of the *xgettext* tool:

```
xgettext -a --files-from=POTFILES.in -o PACKAGE.pot
```

where *-a* is an optional command to instruct the *xgettext* tool to extract all marked strings from the source code files that are listed in the input file *POTFILES.in*. The *--files-from=POTFILES.in* lets the *xgettext* tool to read the names of the input source code files from the *POTFILES.in* file. The *-o* optional command makes the *xgettext* tool to write output to the file *PACKAGE.pot*.

D. Using PO Files

PO files store the translations; every target language should have a single PO file. For instance, a Chinese translation process will translate a PO file into a Chinese (target language) PO file [18]. Normally, a translated PO file contains previous translations provided by the translators. The update of a PO file is done by the *msgmerge* tool. Following is an example of the invocation of the *msgmerge* tool to update the *zh.po* file:

```
msgmerge -u zh0.po PACKAGE.pot --output-file=zh.po
```

where `-u` is to instruct the `msgmerge` tool to update the old version of PO file, `zh0.po`, into a new version `zh.po` in reference to the new version of file `PACKAGE.pot`.

E. Generating MO files

For each PO file (in corresponding to each language), the `i18n` framework will generate one MO file via the `msgfmt` tool. Below is an example that shows how the `msgfmt` tool is invoked:

```
msgfmt -c zh.po --output-file=zh.mo
```

where `-c` instructs the `msgfmt` tool to perform a check on language dependent format strings, contents of the source file header entry and conflicts between domain directives and the `--output-file` option. As the generated `zh.mo` file is a binary file, it is ready for the programming tools (compiler and linker) to use in generating a run-time link library such as the dynamic link library `zh.dll`. It should be pointed out that the `i18n` framework described above is only necessary for software package maintainers or developers. End users do not have to perform any of the tasks showed in Fig. 1 and Fig. 2.

VI. CONCLUSION AND FUTURE WORKS

The `i18n` in an archive retrieval system does not mean that information is simply translated via externalizing its user interface. It is a process that demands for a collaborative effort from managers, designers, programmers, translators and it depends on users' feedback for further development. Every software, module or component needs `i18n`, but `i18n` is not a simply a software translation. Translators may not be able to choose the best phrase in the target language for any text that may possibly be seen by an external user, that is, error messages, help messages and the like. Archive translation software should be concurrently carried out within the whole software development cycle so that translators are able to work within the context about the software and the project. Software that supports Unicode is not necessarily an `i18n` system or product. Unicode is a coded character set. Only characters or parts of characters are encoded, but there is no information about language, locale and font. If a software product can support Unicode, it only recognizes single characters. Unicode is for supporting languages around the world, but it is not a panacea for `i18n`.

Further works are needed to (1) implement a language code interpreter, (2) increase the vocabulary in the indexing database and (3) test and validate the reliability of the system tools.

ACKNOWLEDGMENT

The work carried out in this paper is partially funded by the China National Science Funding Council No. 60873208.

REFERENCES

[1] Wikipedia, "Cloud Computing". The free encyclopedia http://en.wikipedia.org/wiki/Cloud_computing. Retrieved November 2010.

[2] T. Claburn, "Google's "Gov Cloud" Wins \$7.2 Million Los Angeles Contract". The Internet On-line Resources: Informationweek.com <http://www.informationweek.com/news/services/saas/showArticle.jhtml?articleID=221100129>. Retrieved August 2010.

[3] NIST, "Long Term Knowledge Retention (LTKR): Archival and Representation Standards", Gaithersburg, MD 20899, March 2006.

[4] B. Swanson, "The Coming Exaflood", *The Wall Street Journal*, <http://online.wsj.com/article/SB116925820512582318.html>. Retrieved August 2010.

[5] Z. X. Zhao and L. Z. Zhao, 2008, "Small-world phenomenon: toward an analytical model for data exchange in Product Lifecycle Management", *International Journal of Internet Manufacturing and Services*, Vol. 1, No. 3, pp. 213-230.

[6] The ACC Localization Advisory Board, "ACC Globalization Internationalization Localization", Austin Community College, <http://www.austincc.edu/techcert/accGIL.html>. Retrieved March 2009.

[7] U. Drepper, J. Meyering, F. Pinard, and B. Haible,, "Native Language Support Library and Tools, GNU Gettext Tools, Version 0.14.4 (Edition 0.14.4, 8 March 2005)", GNU Project - Free Software Foundation (FSF), Free Software Foundation, Inc., <http://www.gnu.org/software/gettext/>. Retrieved August 2009.

[8] M. Perrow, "Editor's Notes: Better Software Require A Better Process", *The Rational Edge*, January, 2001, pp. 1-2.

[9] J. H. Dunning, "Recent developments in research on multinational enterprises: an economist's view". In Lars-Gunnar Mattson and Finn Wiedersheim-Paul (eds), *Recent Research on the Internationalization of Business*. Proceedings from the Annual Meeting of the European International Business Association, Uppsala, Sweden, 14-17 December 1977. Uppsala: Uppsala University, pp. 1-16.

[10] J. H. Dunning, "Location and the multinational enterprise. A neglected factor", *Journal of International Business Studies*, Vol.29, No.1, 1998, pp. 45-66.

[11] A. Heinrich, "Internationalisation, market structures and enterprise behaviour. The Janus-faced Russian gas monopoly Gazprom". In Kari Liuhto (ed.), *East Goes West. The Internationalization of Eastern Enterprises*. Lappeenranta: University of Technology, 2001, pp. 51-87.

[12] J. L. Nhampossa, and P. Nielsen, "Experiences of internationalizing Information Systems: The challenge of standardization", Presentation at the Oslo/Cambridge IS Workshop at University of Cambridge, UK. http://www.hisp.info/confluence/download/attachments/2374/cambridge_2004_nielsen_nhampossa.pdf? Retrieved December 2007.

[13] A. Vine, "All Things International, only Some of Them Software", The I18n G. A. L, Sun Microsystems, Inc., <http://blogs.sun.com/i18ngal/>. Retrieved September 2008.

[14] S. Kleinfeldt, M. Guiney, J. K. Miller, and M. Barnes, "Design methodology management", *Proceedings of the IEEE*, Vol.82, No.2, 1994, pp. 231-250.

[15] R. A. Baldwin, and M. J. Chung, "A formal approach to managing design processes", *Computer*, IEEE Computer Society, February, 1995, pp. 54-63.

[16] Z. X. Zhao, "A methodology management approach to computerised process planning", *International Journal of Computer Integrated Manufacturing*. Vol.10, Nos1-4, 1997, pp. 83-91.

[17] K. W. Fiduk, S. Kleinfeldt,, M. Kosarchyn, and E. B. Perez, "Design Methodology Management - A CAD Framework Initiative Perspective", *Proceedings of the 27th ACM/IEEE conference on Design automation*, 1991, pp. 278-283.

[18] ISO639 Joint Advisory Committee, "Codes for the Representation of Names of Languages", ISO639.2, November 2006, http://www.loc.gov/standards/iso639-2/php/English_list.php. Retrieved March 2009.

[19] ActivePerl, "ActivePerl: Version5.8.8.817", ActiveState Software Inc. <http://www.activestate.com/Products/ActivePerl/>. Retrieved October 2010.

An Architecture of Virtual Desktop Cloud: Design and Implementation

Zhan-wei Liu Tong-rang Fan Zheng-xu Zhao

School of Information Science and Technology

Shijiazhuang Tiedao University

Hebei, Shijiazhuang, P R China

e-mails: liuzhanwei@hotmail.com fantr@sjzri.edu.cn zhaozhengxu@staff.stdu.edu.cn

Abstract—This paper starts with an evaluation over the virtual desktop cloud technology and its applications in business and forecasts its development in the security and reliability of information systems. It then proposes an architecture of virtual desktop cloud based on the X86 platform. It finally presents the implementation of the architecture and describes how the architecture can significantly reduce the maintenance costs and upgrading cycle of computing systems and facilities.

Keywords- virtual desktop cloud; sever virtualization; desktop virtualization

I. INTRODUCTION

With the development of computer technology and Internet technology, cloud computing is emerging as a web-based process by which shared resources, software, and information are provided to computers and other devices such as smart phones on demand over the Internet [1]. As terminal users increase rapidly in numbers, the demands for both computing resources and shared information are becoming diverse, which make the desktop management extremely complicated and expensive in system installation and maintenance [2].

Conventional IT systems are developed from a centralization paradigm with distributed PCs for independent users. Those systems have its advantages of customized desktop environments, but are difficult to maintain and are vulnerable to errors, risk and disasters. The desktop cloud system described in this paper enhances the security of the desktop environments. It adopts a centralized computation, but relies on the architecture to realize mobile and remote user applications, which represents so-call Private Cloud. The architecture of this system is explained below.

The Teaching and Research Center within the School of Information Science and Technology at the Shijiazhuang Tiedao University has more than 12 computing laboratories, with software and hardware being distributed in different floors and buildings. This computing facility can accommodate about 800 terminal users and is mainly equipped for meeting the educational demands such as postgraduate and undergraduate courses, tutorials and projects. The facility is incorporated with various modern computing technologies including .NET, Linux, JAVA, Database, 2D/3D engineering, to name only a few. In the early days, those computing laboratories relied merely on system recovery cards, isolation units and recoverable network devices to divide the computing resources

(including software and hardware) into different sub-systems and sections. Tailored for different teaching courses, sectional systems had to be set and configured to meet the specific requirements.

In such a scenario, it means that every computing laboratory must install a number of different system configurations for different demands, which had made the overall facility complex and complicated, leading to a significant of management and maintenance costs. At the same time, due to the fast development of hardware and software, the lifecycle of terminal PCs are becoming shorter (upgraded and renewed for every 4 or 5 years), this has added on a large annual cost over the maintenance of the facility. On the contrary, as all other large organizations [2, 3, 4], the budget for purchasing and maintaining computing facilities including PCs at Shijiazhuang Tiedao University has been tightly controlled and for some projects it has reduced, which has inevitably created a dilemma between the system upgrading and the management costs.

In the summer of 2010, the Institution started the project called virtual desktop cloud to bring the state-of-the-art technology and solutions to the aforementioned problems. The project, via virtual servers, is to integrate a large numbers of dedicated servers so that an optimize hardware configuration can be established to reduce the computing resources. Since the so-called Virtual Desktop Infrastructure (VDI) has already been widely implemented on the campus, computing resources previously in the user's desktops can be integrated from the distributed PCs into the data center through remote desktop protocol. The merits of such strategy are (1) to extend the life of PCs, (2) to improve the customer service quality and (3) to generate fast and repaid responses to the customer's requests.

The implementation shows that the virtual desktop technology has enabled PC customer's access their own virtual desktop through any network port or equipment and the computer lifecycle has been significantly extended. Because all virtual desktops have been integrated in the data center of cloud computing instead of being distributed over a population of single PCs, very little resource is needed for installing patch programs or security update in the individual PCs. All previously existing system management and maintenance tasks can be carried out in a cloud computing architecture (within the data center), which provided a much more flexibility and robust environment for students and enabled them to access their own personal desktop anywhere (and with much more mobility).

With the aforementioned application example, this paper presents a virtual desktop cloud and demonstrates its design and implementation as a web-based process by which shared resources, software, and information are provided to computers and other user-centered devices, proposing an architecture that is different from conventional IT centralization infrastructure [5, 6, 7].

In the text to follow, Section II describes the proposed system structure and Section III illustrates the work scheme of the system and Section IV demonstrates the functionality of the architecture. Section V provides a comparative study to validate and justify the architecture against the IT centralization paradigm. It also provides a brief description of user experiences in terms of the design flexibility and customization for different applications. Section VI shows conclusion and future work.

II. SYSTEM STRUCTURE

Virtual desktop cloud technology integrates comprehensive virtualization technologies toward servers, OS's, desktops, thin client, remote link protocol, and so on [1]. Virtual desktop cloud solution differs from others in that it can deliver personal desktops to customers by using single mirrors. This solution simplifies desktop management and improves the service quality, by which system administrators are able to choose distributing systems more flexibly. This enables the system to distribute desktops to individual computing laboratories, remote libraries, staff offices and student dormitories.

With the current installation and configuration of the servers in the Teaching and Research Center in the School of Information Science and Technology at the Shijiazhuang Tiedao University, two virtual partitions are used where desktop virtualization platforms and DVC [1, 2] (distant visual cluster) software are involved. The first one has the capacity of supporting 400 desktop platforms, which is built through utilizing desktop virtualization platforms and enterprise-level servers and disk arrays. The second one is a distant visual cluster system built with high-performance computers with distant visual cluster software, which is mainly for graphical design and research projects on the campus.

A. Architecture of VisualView Software

The desktop virtual platform has an architecture that includes an end-to-end solution and can deliver desktop applications in the form of managed services [1]. This architecture is shown in Fig.1. The virtualization platform provides a highly scalable, highly reliable and stable platform for running virtual desktop applications, which has the continuity of services and disaster recovery functions to protect the user information and the desktop data [2]. The platform provides a guarantee for desktop virtualization, and is inexpensive and simple like traditional solutions. The management center can completely control and check clusters, host computers, virtual computers, memory, network connections and other key factors within the virtual basic architecture [3, 4].

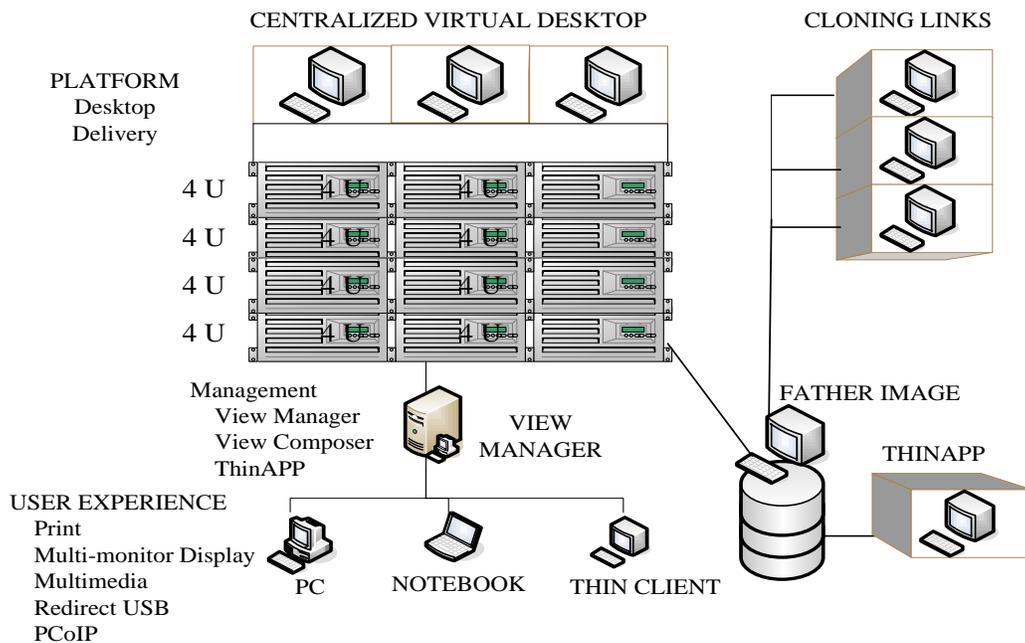


Figure 1. System Frame of Virtual Desktop Cloud.

The virtual manager software can in theory manage thousands of virtual desktops from a single memory image for controllers, which simplifies desktop management, allocation and deployment. At the same time, terminal users can access virtual desktops safely and easily through the Software-View Manager [4]. In an educational establishment, students are able to establish rapidly a desktop image shared with a single up-level image in virtual disks through virtual composers. The virtual composer may segment the user data and configuration for independent management, so it will not affect the user data and configuration whenever repairing and updating desktop linked with up level images need to be carried out [5]. The Virtual ThinApp simplifies the management and distribution of applications, which can rapidly dispose the applications to users and avoid any data transfer conflicts [2, 3, 4].

For software applications, the virtual manager is able to manage as many virtual desktops from a single memory image for controllers as they are needed. This arrangement simplifies desktop management, allocation and deployment. At the same time, terminal users can access virtual desktops safely and easily through the Software-View Manager [4]. Students are able to establish rapidly a desktop image shared with a single up-level image in virtual disks through virtual composers. The virtual composer may segment the user data and configuration for independent management, so it will

not influence the user data and configuration whenever repairing and updating desktop linked with up level images need to be carried out [5]. The Virtual ThinApp enhances the functionality of management and simplifies the distribution of applications. This architecture can therefore rapidly dispose the applications to users and avoid any data transfer conflicts.

B. Architecture of Distant Visual Cluster Systems

The distant visual cluster DVC [6] adopts a C/S architecture including server-end and client-end. The server-end realizes OpenGL to accelerate the rendering cycles of user applications and it compresses the rendering image for a fast data transmission. For the client-end, it receives the compressed images and extracts to display on the monitors. The detailed architecture of DVC for these functions is shown in Fig 2. The sever-end is connected with the client-end through specific protocols. Virtual displays are virtualized by software in the server which directly sends 3D graphic operation commands to the 3D acceleration graphic card in the server-end. Test shows that this arrangement can significantly improve the rendering efficiency via utilizing accelerated rendering of 3D drivers. Events of mouse and keyboard can be sent to the client-end through the protocols in order to run applications in the server-end. These specific processes are list below.

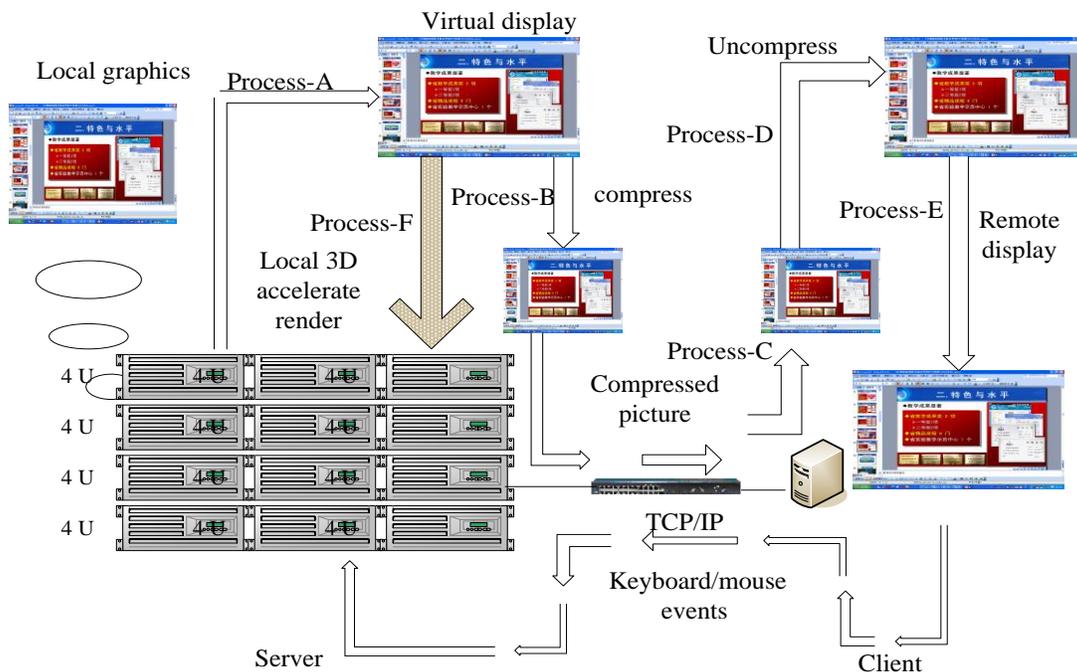


Figure 2. Architecture of DVC system.

1) *Process-A*: The local server-end utilizes the virtual display and the process-F to realize the hardware and the 3D accelerated rendering process in the server-end and it thus forms a 2D rendering graph in virtual displays which is of high-resolution and high-capacity [6].

2) *Process-B*: This is to compress the high-resolution and high-capacity rendering graph to small-capacity graphs that are suitable to transmit data in the network. It utilizes the image compression technology with high compression ratio.

3) *Process-C*: This process transmits the compressed graphs in the server-end to the distant client-end through internet.

4) *Process-D*: This decompresses the graphs and the graphs are received from the distant client-end.

5) *Process-E*: The rendering 2D graph and displaying in client-end are processed in this phase.

6) *Process-F*: The virtual displays send the accelerated rendering commands to the 3D accelerated hardware.

In the transmission between the sever-end and client-end, the arrow represents the transmission process of the mouse and keyboard events. From the above processes, it is shown that the geometric model of the rendering section with huge amount of calculation running on high speed graphic card and the 2D graphic compressed data packed is extracted and displayed on the monitor.

Data transmitted in the internet includes rendering graph and mouse and keyboard event instead of 3D geometric model data. This as a result substantially reduces the amount of data transmitting in the networks and decreases the dependence on networks, thus it improves the operations for graphic applications. This is one of the obvious advantages in employing DVC to realize distant 3D accelerated rendering for applications.

III. DESCRIPTION OF SCHEME

A. Configuration of Virtual desktop Cloud Server

The configuration of the server used for virtual desktop cloud is listed in Table I. The server is a standard 4U rack which is an enterprise server whose main configuration can be simply summarized as below.

TABLE I. CONFIGURATION LIST OF VIRTUAL DESKTOP CLOUD SERVER

Standard 4U rack enterprise server	
Processor	Opteron 6134 (2.3GHz, 8 cores) *4
Memory	128GB ECC DDR3 1333 Registered memory; supporting advanced management function
Hard disk	146GB, hot-swap 2.5 inch SAS *2
RAID card	512MB SAS RAID card, with battery, supporting RAID0/1/5/6/10/50/60
NIC	2 port gigabit-NIC with 5 gigabit network interface
HBA card	single port 8Gb PCI-E optical fiber HBA card *2
power	1+1 redundancy hot-swap power module
fan	Front 2+1 redundancy hot-swap fan
Management software	Integrated IPMI, iKVM, virtual media; distant management software, condition monitoring software for equipments; information management and distant access software; backup and recovery software for sever systems

The processor is an Opteron 6134 with 8 cores; the memory capacity is 128GB; the hard disk is 2 SAS hard disks with 146GB capacity.

B. Memory Configuration for Virtual Desktop Cloud

The memory configuration for the virtual desktop cloud has three parts which are as follows.

1) *Requirements of memory capacity* For the proposed architecture, each user is allocated with a 20GB disk. This capacity is determined for both systems described earlier. A 30GB disk capacity is used for data storage according to literature [6]. In total it needs 20TB disk capacity for 400 users.

2) *Configuration of memory capacity* Using a fiber channel storage array, a main cabinet and an extended cabinet, the configuration and distribution are list below.

a) The 16 600GB/block disks provide a 8.4TB bare space in which one block is used as a hot backup and the other 15 bocks constitute the RAID5 module.

b) The 14 1TB/block SATA disks is able to provide 12TB bare disk space in which one block is used as a hot backup and the others form the RAID5 module.

c) The system space is placed on the FC disks and the data space is placed on the SATA disks.

3) *Storage performance* The fiber channel storage array has 8Gbps channels and is constructed to provide a transmission capacity of 20000 I/O per second which is to meet the 400 I/O requirements designed for virtual desktop applications.

C. Planning of Network

The communication networks among severs use an independent 1000Mbps Ethernet, which is intended for the data transmission between the virtual servers. The design details of the network are listed in Table II.

TABLE II. PLANNING OF NETWORK

Equipment type	Number	Detailed configuration
HPC platform case	1	Independently supporting 5 nodes, 2 gigabit exchange module, 2 management modules equipped with 2 2000W redundancy power modules, supporting UPS modules, memory modules and management PC modules
Workstation node	5	2*XeonE5620 /2*SAS2.5 inch 10K147G/ 24G/Nvidia Quatra FX3800 professional graphic cards
PC module	1	160GB hard-disk, processor and 1G memory integrated system with mouse, keyboard and monitor
DVC	1	Distant virtual workstation (cluster system)

This part describes the communication networks between the servers and the memories. The networks employ a two-

link redundancy mechanism with two 8Gbps which are capable of preventing all applications from sudden shutdowns that may be caused by faulty devices or communication units.

D. Configuration of Distant Visual Clusters

The distant visual cluster includes one high-performance computer platform, five workstation nodes, one PC module, one UPS module and the corresponding virtual workstation cluster software.

Table III is the details for such a configuration. The communication network between server and memory has two sections, one has an 8Gbps FC and the other includes two link redundancy mechanisms. The communication network between server and server is a 1000Mbps Ethernet.

TABLE III. CONFIGURATION OF DISTANT VIRTUAL CLUSTER

Type	Design
Communication network between server and memory	A. 8Gbps FC B. two link redundancy mechanisms
Communication network between server and server	1000Mbps Ethernet

IV. SYSTEM FUNCTIONS

The system functions so far have fallen into two categories in terms of its structure design. One is based on the architecture and the other is on the cluster applications.

A. Architecture of Virtual Desktop

The virtual desktop employs an optimized cloud computing platform which serves as its underlying architecture. This platform provides efficient server virtual functions. As a background support, it has following five features.

1) Extensibility

Each management unit supports up to 1000 virtual machines, which makes it suitable for a large deployment of virtual desktops. By using vMotion [7], the system is made more efficient and faster than conventional IT infrastructures and the migration time can be shortened significantly. Depending on service demands and the priority to compress and add desktop applications, the server resource is able to distribute in a dynamic module.

2) High performance

vSphere, equipped with high performance, is able to provide a fast and stable platform for the virtual desktop applications and to obtain an optimal status of the servers and virtual machines by using the monitoring platform [7].

3) Optimum density

With the increasing density of virtual desktops, there are 16 to 20 virtual desktops per core, this can increase the numbers of the supporting machines in each sever.

4) High-availability and business connectivity

vSphere optimizes the workload of the desktops. The performance improves because of the reduction of the memory exchange.

5) Rapid disaster recovery

Both the data recovery technology and the vMotion [7] technology are able to provide the safety of the virtual desktop platforms.

The system administrators are able to use the virtual desktops as a central controlling node. This node supports terminal users for safe and flexible accesses to the virtual desktops and is able to deliver the desktops in a style which is called the managed security service model. The virtual management software possesses an expansibility and reliability and utilizes a management interface. This interface is formed for the Web services to create and update the desktop images, to manage the user data, to implement the global strategy and to manage and monitor as many virtual desktops as it is currently needed (about 1000) simultaneously.

B. Distant Visual Cluster Software

The 2D or 3D software is needed for routine teaching and research needs. Those software systems are used to meet visual demands for processing the exchanged data. Base on a number of tests and experimental tries, the solution of DVC has solved the aforementioned problems. The configuration of the system is described below.

1) Improving the security of data.

Distant users are able to access and operate the corresponding applications; there is no other data to transmit to the distant users except for the distant desktops which ensures the safety of the user data.

2) Improving hardware utilization.

By using the DVC software, one workstation can support a number of different users at the same time, thus improving the rate of facility utilization.

3) Improving efficiency.

Through the DVC software, staff and students are able to work anywhere and participate in a virtual environment that is efficient and effective.

4) Reducing management cost.

With the DVC software, the hardware and software can be integrated seamlessly, which effectively reduces the management costs, improves the operational efficiency and extends the life span of the systems that are involved.

V. COMPARATIVE ANALYSIS OF SCHEME

The above design and implementation has been used in a one year teaching and research environment as a testing period and the user experience indicates that the virtual desktop cloud solution is such an implementation that is able to support in an enterprise level the distant dynamic access of desktop systems and the unified managed technology of data centers. In comparison to the tradition IT systems, it is a new module that is based on servers and thin client modules so that system administrators and users can take the advantages of the two modules simultaneously.

The results show that all desktop virtual machines are trusted and uniformly managed in the data center. The users

can have the same user experience as but better results than with the traditional IT systems via the thin client, the similar equipment in LAN or distant access. Especially under the virtual desktop cloud architecture, the openness and zero-touch of the cloud computing basic architecture can be realized. In the transaction process of shift toward the cloud computing from the traditional paradigm, robust data protection and full utilization of resource can be achieved.

The vSphere can realize a free migration between the servers and the virtual machines. This will help in realizing an automatic detection of fault. It can also obtain the distributed resource allocation to realize a balanced workload among applications. The vMotion can help in achieving the real-time migration between running servers, obtaining zero-shutdown characters, which is able to enhance the availability of the servers and to increase data security.

It is also justified in the testing period that the virtual desktop architecture provides a safe and reliable data storage center, protecting the users from data loss, virus entry and other online hazards. A convenient and fast 'cloud server' can reduce the workload compared to the old daily maintenance work. It is verified that the system administrators are able to carry out easily the maintenance work including maintaining hardware, installing and updating software, preventing virus and network attacks. Finally, users only need to type their address or login details to access the system and carry out the work exactly the same manner as on PCs. Cloud computing provides almost infinite space for storing and managing the data and provides the most so far capability for completing large applications.

However, as for information security, user experience, existing bandwidth, product type choosing and allocation respects, the virtual desktop planning still face many technical and commercial challenges. For security and privacy of personal data, identity authentication and data backup should be enforced to ensure the high reliability and availability of data in the implementation phase. Since various mobile storage devices are being used today, printing and transmission of streaming media data may restrict the users to access the virtual desktops and their application data.

VI. CONCLUSION AND FUTURE WORKS

In order to realize IaaS (Infrastructure as a Service), PaaS (Platform as a Service), SaaS (Software as a Service) of computing resource and to simplify terminal units, system resources should be integrated and managed in cloud units to improve the system efficiency and enhance the service quality. The architecture of virtual desktop cloud makes it possible for different computer systems and computing resources to be managed centrally in the data center and distributed through the network in service modules. This makes it possible meet the user's demands. By integrated

management of computing resources, idle computing units and storage can be reduced substantially.

Future works are needed to (1) carry out further study toward empirical validation of the system so that the proposed architecture can be justified for its claimed functions; (2) the development process needs to be further validated and more user experiences need to be gathered to examines the merits and failures of the design and implementation, especially the design flexibility for new required features and the customizing ability for different applications.

VII. ACKNOWLEDGMENT

The work carried out in this paper is partially funded by the China National Science Funding Council. Funding number is 60873208.

REFERENCES

- [1] Wikipedia. http://en.wikipedia.org/wiki/Cloud_computing. Retrieved September 2010
- [2] China Cloud computing net. <http://www.cloudcomputing-china.cn/Article/ShowArticle.asp?ArticleID=1>. Retrieved October 2010.
- [3] VMware official site. <http://www.vmware.com/cn/solutions/virtualization-management/>. Retrieved July 2010.
- [4] Dawning official site. <http://www.dawning.com.cn>. Retrieved August 2010.
- [5] J. Varia. "Cloud architectures- Amazon web services". ACM Monthly Tech Talk , <http://acmbangalore.org/events/monthly-talk/may-2008--cloud-architectures---amazon-web-services.html>, Retrieved August 2010.
- [6] I. Foster, C. Kesselman, and S. Tuecke, "The anatomy of the grid enabling scalable virtual organizations". International Journal of High Performance Computing Applications. Vol. 15, No.3, August 2001, pp. 200-222
- [7] F. Berman, G. Fox, and T. Hey. "The grid: past, present, and future". "Grid Computing: Making the Global Infrastructure a Reality". John Wiley & Sons, Ltd, 2003, pp. 9-50.
- [8] Top 500 supercomputing sites. <http://www.top500.org/>. Retrieved July 2010.
- [9] A. S. Szalay, P. Kunszt, A. Thakar, J. Gray, D. Slutz, and R. J. Brunner. "Designing and mining multi-terabyte astronomy archives: The Sloan Digital Sky Survey". SIGMOD International Conference on Management of Data Proceedings of the 2000 ACM SIGMOD international conference on Management of data. ACM, Vol. 29, No.2, 2000, pp. 451-462
- [10] L. A. Barroso, J. Dean, and U. H-Izle, "Web search for a planet: The Google cluster architecture". IEEE Micro, Mar/Apr, Vol. 23, No. 2, 2003, pp. 22 - 28.
- [11] Google tops translation ranking[N]. News@Nature, <http://www.nature.com/news/2006/061106/full/news061106-6.html>, Nov. 6, 2006.
- [12] R. E. Bryant, "Data-Intensive supercomputing: the case for DISC". CMU Technical Report CMU-CS-07-128. May 10, 2007, pp. 1-10.
- [13] S. Ghemawat, H. Gobioff, and P. Leung, "The Google file system". Proceedings of the nineteenth ACM symposium on Operating systems principles. Oct. 2003, pp. 108-114.

A Study of Dragon-lab Federal Experiment Cloud and Network Contest

Tongrang Fan

School of Information Science and
Technology ,
Shijiazhuang Tiedao University
Shijiazhuang, China
Fantr2009@126.com

Zhanwei Liu

School of Information Science and
Technology ,
Shijiazhuang Tiedao University
Shijiazhuang, China
Liuzhw @stdu.edu.cn

Liping Niu

School of Information Science and
Technology ,
Shijiazhuang Tiedao University
Shijiazhuang, China
niuliping666@163.com

Abstract—Through the research and analysis on the contests of the IT organization management mode, operation platform and the trend of Network Contests (NC) development , the paper explores the issue about Dragon-lab federal experiment cloud and network contest environment construction, furthermore, it proposes a comprehensive solution for network contests layered management model based on the cloud computing. The solution has the following novel features: (i) a hierarchical level of contest management model based on cloud computing resources; (ii) design for network framework based on the Dragon-lab federal experimental cloud; (iii) service operation mode in Dragon-lab experiment cloud; (iv) scheduling strategy for the contest management platform. The advantages of scheme lie in: (i) eliminating the limitations of contestants, physical location, match point location and hardware/software resource space in the maximum; (ii) promoting the scheduling and sharing of device resource among the different NC areas; (iii) reducing the NC cost and improving the efficient utilization of the tournament game equipment. Finally, the solution had been verified, in Dragon-lab experiment cloud, by successfully hosting network skills competition in four NC regional sites.

Keywords- cloud computing; dragon-lab federal; Network Contest; experiment cloud; contest management model

I. INTRODUCTION

The contest based on computer use and development started from 1970. Since the Texas A&M university held the first games, this new discovery and the way of developing top students majoring in computer science, responded positively by some U.S. and Canadian universities immediately[1]. In 1977, the international collegiate programming contest (ACM/ICPC) organized by ACM has become a classic event-an annual session of multinational international computer programs.

With the booming of Internet technology and rapid development, contests about network equipment use skills have appeared in recent ten years. Among those the most famous are: ICTF UCSB Contest[2], Collegiate Cyber Defense Competition[3], NTU Network Security Competition[4], and China university NOC description activity network security competition, network skills contest sponsored by education ministry[5], etc. Competition contents of Network Contests mainly focused on the design and implementation of network security solution in LAN of the enterprises, campus and government which emphasizes

network attack and defense technology. The purpose of NC is to test contestants' capabilities in understanding and manipulating Internet infrastructure and business information system security. This kind of competition demands large amount of network equipment, meanwhile, the management of the contest is a little complex.

The competition management of computer program mostly use the following systems: Programming Contest Control (PC2) and Mooshak system[6]. Both are based on C/S structure of competition system which only support single regional sites competition, while do not support division online competitive[7][8]. This system cannot directly support network tournament organization management.

In recent years, with the increasing number of participants in Network Contests, the competition modes are also developing toward the trend of multi participant areas and multi participant spots. Obviously, the existing local contest organization management cannot meet the requirement of the of network tournament's development. Therefore, how to eliminate the limitations of physical location of the traditional tournament mode, how to take the advantage of network environment in maximum in order to satisfy the demands of the network contests which includes depending largely on network equipment, large quantities, variety types, complex requirements and high requirement of hardware and software environment, becomes the major problem of our study. In the process of research, we propose the Dragon-lab federal experimental environment and construct experimental cloud network to solve the problem of effective use of network equipment; Using cloud computing technology in order to develop contests management system based on cloud service environment, promoting the scheduling and sharing of device resource among different NC areas, and verifying the effective of this idea by holding a four areas of the joint tournament. The value of this study lies in expanding the application field of cloud computing.

The use of cloud computing in contests has not been reported yet[9~13]. The construction of contests is developing rapidly, therefore, the unity resource platform and regulation rules will be benefit to network contests. For some contest sponsors, the cost of building contests computing center is too high which does not match with the fast development of contests and diversification of services. The cloud computing mode supply contests sponsor with

appropriate schemes, using cloud computing can coordinate the network infrastructure, network equipment and the tasks of data center based on contests. This mode can effectively reduce the cost and the work of maintain, upgrade and update equipment can be done less.

Maximize the resources sharing. Using the stronger management mechanism, automation deployment and high level of virtuality function of cloud computing techniques, to realize the maximization of network virtual environment resource sharing and co-work.

II. THE BASIS OF CLOUD COMPUTING

This section illustrates the key technologies currently used in cloud computing which lay a foundation of contest organization management based on cloud computing that proposed in the next section.

A. The types of cloud computing service

The types of cloud computing service have three mainstream types according to the types of classification approach[10]: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS).

Infrastructure as a service (IaaS) is the basic facilitate. Service providers offer servers, storage and other hardware, so users' application system and software can be managed by the Service providers.

B. The application of cloud computing

Cloud computing, based on virtualization and web, providing services such as basic architecture, platform, and software, integrating the large-scale extensible computation, storage, data, and application of distributed computing resources to carry on the work together. As a brand new Internet application mode, this paper mainly discusses the following points.

(1) Virtualization technology

Virtualization technology provides an effective solution to resource management in cloud computing. By sealing the service in virtual machine and rejecting to each physical server, virtualization technology can remap the virtual and physical resource according to the change of loading, as a result, the balance load of the whole system can be realized.

(2) Security technology

It includes the following six areas: (i) The security about access control of cloud. (ii) Data security and privacy protection. (iii) Audit Security. (iv) Storage security. (v) Defense security. (vi) Safety laws and regulations.

(3) Storage technology

The data storage layer of cloud connects the different types of storage devices to achieve the goal of unified management of massive data and centralized management of storage devices, status monitoring, and dynamic capacity expansion. In essence, it is a service-oriented distributed storage systems to meet the need of performance and storage capacity requirements under the conditions of multi-user.

(4) Scheduling and allocating technology

The scheduling and allocating in cloud computing supports three levels management: (i) It specifies how much processing power per core in a host should be assigned for

each virtual machine. (ii) The virtual machine allocates an available amount of processing power to independent unit of the task. (iii) It must decide to choose which data center for the user and use the strategies of resources management and model of cost about access in the data center for cloud services.

III. NETWORK CONTEST ORGANIZATIONAL MANAGEMENT MODE

The whole contest is generally managed by three institutions. Its management system adopts B/S structure, and running on the Dragon-lab federal experiment cloud network platform which based on CERNET, supporting fast, efficient communication command and control, so it can avoid congestion caused by all the participants access to a site simultaneously.

The three levels of management institutions are: organizing committee, executive committee and sub-regional contests. (i) The organizing committee is responsible for managing the affairs of information center about contest. (ii) The executive committee is responsible for the management of the business center in sub-regional contests. (iii) The sub-regional contest is only responsible for the competition in the LAN and monitoring the competition, checking the eligibility and marking the features for contestants online. The number and the position of the centers for information and business management can be dynamically arranged based on the actual demand.

The service target of the contest management system are contestants, judges in review committee, staff of the centers for business in executive committee, staff of the centers for information management in contest organizing committee and visitors who have not registered. The jury, the organizing committee and the executive committee are related to the management of organizations, in the management system for contest, they are called the judges, staff of the centers for information management, staff of the centers for business separately.

The management system of the network contest is divided into two corresponding subsystems which are the management systems of the business and information. The management system of business manage sub-regional contest and be responsible for competition device resource scheduling and judging sub-regional contest. The management system of information includes events proposition and event information openness. By using the strategy of cloud computing, the contest management platform can resolve the problem of space limitation in the contest user, location, hardware and software resources. The model of hierarchical architecture about contest was shown as Fig. 1.

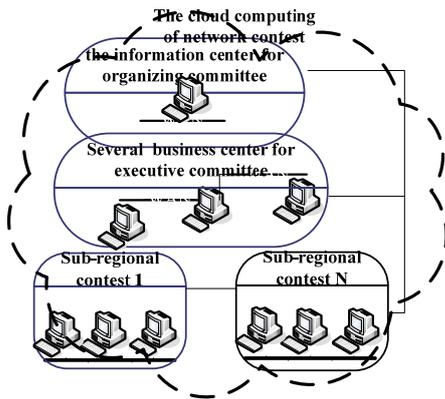


Figure 1. Example of a model of hierarchical architecture about contest

IV. CONTESTS MANAGEMENT BASED ON DRAGON-LAB FEDERAL EXPERIMENT CLOUD

A. Contests management based on Dragon-lab federal experiment cloud

Federal is a unique form of cooperation. It mainly consists of two meanings: autonomy and cooperation. Autonomy means that the resources of the federal are invested, maintained and cooperated by members autonomously and voluntarily. The federal of network infrastructure is a loose coupling form, resources invested by members may not be stable, but with a certain level, the effects of random will gradually decrease, while the advantage is obvious, that is, low operating costs and strong sustainable development.

Dragon-lab[14] is the research platform of CNGI-CERNET2 which is a large-scale IPv6 Internet's backbone network belongs to China Education and Research Network.

It is the only large-scale next-generation Internet technology cooperative research platform that based on federal structure.

B. Key characteristics of Dragon-lab experimental cloud network

The network is run in multi space-time environment. Dragon-lab will integrate all kinds of network application environments including the center of internet exchanges, international backbone networks, national backbone network, campus network, satellite, wireless, etc, which can record and playback IPv4/IPv6 network flow at any time, and import OSPF, ISIS and BGP routing information from running network. Therefore, through Dragon-lab long time and large space span, real network behavior research and running can be carried out.

Unified and efficient management of network equipment[15].By using Dragon-lab special configuration system, it can manage the existing laboratory equipment unified, and break the original laboratory situation of isolated and inapplicable. By creating a unified experiment platform, and making full use of the original equipment, the local

comprehensive designed network experiment become possible.

Remote visualization experimental configuration. Dragon-lab is a remote laboratory, which can customize the lab environment remotely by using a dedicated client program by means of visualization.

Programmable experiments. Dragon-lab presents an innovative idea about programmable experiments. The experimental repeatability is improved by translating the lab environment definition into an executable test script.

C. Dragon-lab Network Contests management

Network contests management system, running in Dragon-lab federation experiment network cloud platform, contains three parts: the organizing committee contests information center, executive committee business management centers and area sites. Network infrastructure which is used by contests also has three-layer in figure 2. Cloud computing mainly supply the following services: (i) infrastructure service, namely multi-area management which is in charge of the executive committee, supply hardware resource, distribute and adjust those network equipment resource through virtualization technology. (ii) software service, dynamically using resources such as software tool, application system. (iii) data service, cloud computing data supply data sharing, data storage and data recovery.

In the Fig. 2, Dragon-lab federal runs in the China education online, and the core main node is in Beijing in Tsinghua university. The nodes of the first level are provincial node, which is the center of the contests information, called CCI (including data center, contests information management platform, the backbone network equipment, etc.). Secondary nodes are division business center (division data center), administrating the business of the contests and the next level of resources, such as servers, host, switches, routers, network security equipment, and the equipment of agreements and flow analysis, equipped with division competition business management platform. The nodes of the third level are tournament competition points which connected with many kinds of network equipment. Through the above level nodes, scheduling, sharing equipment resources can be realized.

Through login the spots management systems of the business, the user can use the resources of each division management center to take part in the contest. Area business management center is responsible for the division review and management. The contact between the contests spots and the area management center are randomly composed and expanding easily. Business management systems of the contests spots which use cloud computing can improve the expansibility, maintainability and resources utilization of the system, etc.

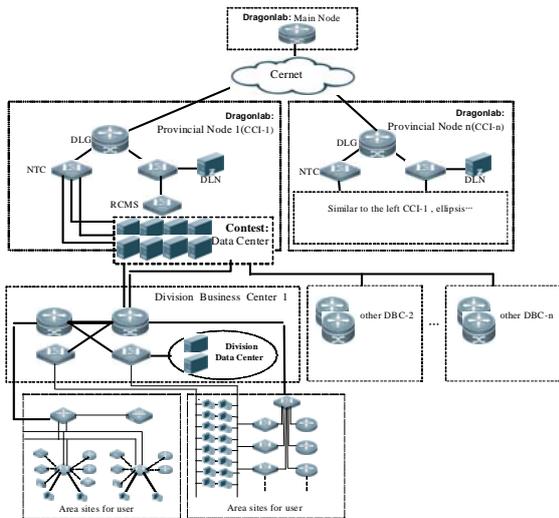


Figure 2. Dragon-lab federal experiment cloud Network infrastructure.

V. SHARING CONTESTS RESOURCE IN ENVIRONMENT CLOUD

The contests resource service running in cloud computing complete are achieving through running in contests/areas data center. This service divide into 5 classes: data discovery, resource allocation, resource storage, resource scheduling and monitoring.

A. network resource running Ways in Dragon-lab experiment cloud

•Cloud computing service agents assign the competition task which is submitted by area users to appropriate area data center. Area data center agent will receive request of all kinds of contests application examples, and then change these requests into cloud task of contests data center, schedule Dragon-lab experiment cloud software/hardware infrastructure resource through virtual machine, complete tasks of real time monitoring, scheduling and resource allocation. The detail flow is shown in Fig. 3.

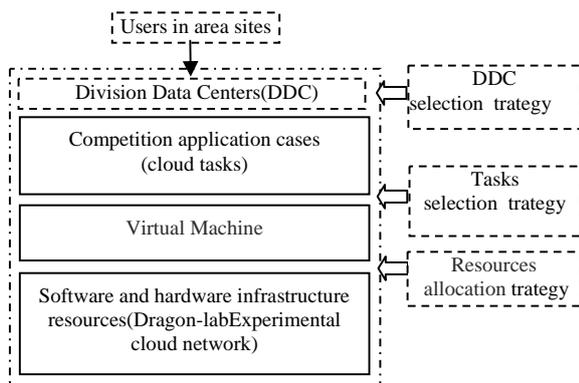


Figure 3. Example of resources scheduling and management in Dragon-lab Experimental cloud

Balance loading of dragon-lab experiment cloud network infrastructure resource is completed through scheduling strategy in Fig. 4. The scheduling strategy mainly includes creating, using, destroying and so on.

New contests data center register service information into cloud service agent. When cloud service agent receive request of area user, it chooses appropriate area data center, schedules and arrange cloud resource task based on user service quality demand, and then take charge of real time coordinate task between user and service.

The monitor center is located at the data center agent of division, in each terminal of network shared device of the division data center, there are configuration of monitor nodes. The monitor nodes are responsible for monitoring the status of computing power resources and usage of the network device, the main information is the property of equipment resources, IP address, the size and usage of CPU and memory, etc real-time information; These information is passed to the monitoring center. The monitoring agent uses the polling monitor strategies, according to the information of the terminal of the monitoring nodes, records the real-time situation of each node periodically and dynamically. Each division data center adopting disaster recovery technology, achieves storage of the important data of network events.

Fulfill distribute computing resource on demand by creating visual machine in the area of data center. Each visual machine is correspond to the resource of network device. Visual technology can make the visual machine remap to physical resource according to the change of load, through packaging services in the visual machine and mapping to the specific network device. The software of visual platform can cut proper size of visual computer pieces (including the kind of device, the number of CPU, the memory size .etc) according to the demand of the contest. The task is running on the visual machine, and data center assign visual resource. By creating, destroying and transferring visual machine periodically, the load balance of cloud network infrastructure resources of Dragon-lab experiment can be reached.

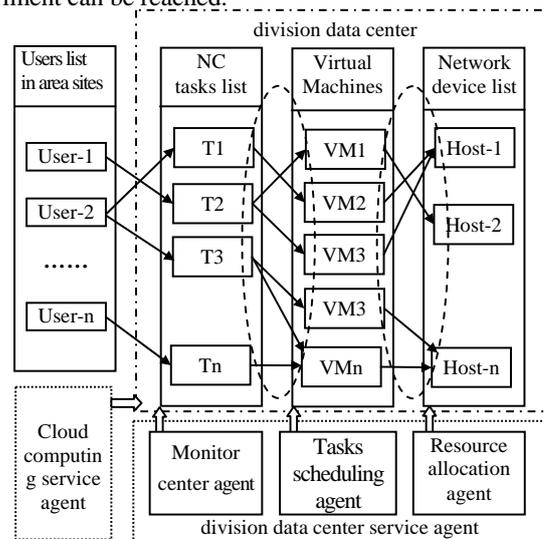


Figure 4. Example of Load balance in Dragon-lab resources

B. The scheduling strategy of test management platform

The strategy executed by contest/area data center service agent includes: monitor strategy, resource distribution and task schedule. It can cluster and divide resource according to K-means algorithm. The resource in the data center can be divided into several clusters according to QoS in order to find the proper resource which matches the task quickly when schedule and distribute resource to data center, by this way fulfill the common QoS reputation of users, then complete this task.

Work scheduling strategy and resource allocation strategy use the way of first come first service, and join the general expectation constraint of the user service quality QoS, and budget limit and time limit, matching the service tasks of the user and the resources of the virtual machine with network equipment. The specific methods are: normalizing the data of the tasks, making the normalized task data and general expectation constraint data as equipment quality, CPU, memory, bandwidth, expenses, at last calculating the minimum distance of equipment resources according to the Euclidean distance equation "(1)" namely the host resources which is the most conformed to the user QoS general expect, and then making the virtual machine resources of executing the tasks and the network equipment resources matched. The scheduling algorithm involves 5 aspects.

$$D(x, y) = \|x - y\| = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

- The selection function of data center: sequence the data registered in the cloud service agent by priority, according to user's QoS demand.
- Resource Clustering: cluster and device the source in data center and sequence it by computing power through K-means algorithm.
- Function of task parameters and classification: set general expectation vector of task QoS, and assign parameters of classification when submit task by user. Creating virtual machine for task execution.
- VM resources parameterized algorithm function: calculating standardized resources parameters on VM host.
- Matching task with resources: In accordance with the classification of task, the task of each category correspond to the general expectations and parameters of the resource vector is matching, computing the Euclidean distance algorithm to achieve the task with the host VM resource binding.

VI. PROVING CASES

The second College Network Skill Contest for Hebei Province in China was held under Dragon-lab federal experiment Cloud Network environment. Contest resources were distributed in four colleges, which are Tsinghua University in Beijing, Shijiazhuang Railway University in Shijiazhuang, Hebei Polytechnic University in Tangshan, and North China Institute of Aerospace in Langfang. Shared network device resources composed of switches, routers and servers supporting IPv6, etc. Using cloud computing service mechanisms for this network contest works, it was very

effective to realize the share of the various types of network experiment resource, and it can also reduce concentrated calls of network devices for contests, the players can join contests nearby, which greatly reduced the cost of contests.

The typical applications in this case, which using test bed for Next Generation Network--Dragon-lab federal experiment platform, had proved a number of network services for IPv6. The content IPv6 address configuration and planning of covers routers, switches and servers, such as: (i) Router configuration: address translation between IPv4 and IPv6, configurations of static routers and OSPFv3 router protocols, ISATAP tunnel configurations, etc. (ii) Switch configuration: switches' IPv6 address configurations, divided VLAN and no-state configuration for subnet getting IP address. (iii) Parts of system application: configuring the ISATAP tunnel on PCs; configuring IIS services on servers which using IPv6 address of servers for access, etc.

VII. CONCLUSION

With the development of cloud computing technology, it will expand a new development space for the contests field. The management and information processing of the contests will gradually migrate to the cloud, which will have a significant and far-reaching impact on the areas of competition. Contest participants can make better use of the information resources and services with the cloud computing services. Contest management agencies can manage the contests resources, organize and operate the contest process much better.

Cloud computing obtained four achievements in the application of network contests:

A. Integrate the resource and improve the service functions of the competition

Through the integration of contests resources, the contests cannot be effected by the geographical location, data processing ability of server etc. The construction of the contests resources need to emphasize the introduction of new technology and the construction of contests contents, meanwhile, emphasize the utilization, retrieval and sustainable development of resources, improving the service consciousness of resource construction, satisfying the effective sharing and utilization of information and resources. Tournament organizers and contests spot who using clouds large-scale server groups, have powerful computation ability and high bandwidth, can rapidly respond to the request of user.

B. Reducing the cost of hardware, providing economical software service

Using cloud computing, most tasks of computing is carried out by cloud end, the computer is just connected to internet. By cloud computing, purchasing cost of hardware and software will be reduced greatly. The demand of terminal equipment is low by using cloud computing. Therefore, cloud computing will be popular in contest field for it can reduce purchasing and maintaining cost for the contest host. Some commonly used software such as online document editing tools, contest software tools can adopt

cloud computing. If the contest areas connect to the service of cloud computing, the cost of software construction and the expense of maintain and update will be reduced. What the users do is just open the explorer, use the cloud computing, perform the contest management, contest staff management and take part in the contest.

C. *Building contest Platform of network competition , convenient for contest and study*

With the gradual development and popularization of Cloud Computing pattern, processing varieties of information about competition will be gradually transplanted to Cloud, which will take a positive influence on network competition. Cloud Computing will have an active influence on building environment of colonial competition and individual autonomy learning and implementing change and sharing information. Users can freely choose related methods, among those the service, resource and platform are provided by Cloud Computing, which provides network competitor and learner abundant resource of network competition and a favorable platform for competing and studying. That will help to launch network competition.

D. *Providing more secure and reliable data center to avoid illegal attack and destruction*

With the virus and hacker prevailing, the security and reliability of data are becoming more and more important. Cloud Computing service is severing for competition which uses the most advanced data center in the world to store data. There is a powerful technological management team managing the committed data, which provides the competition reliable and secure data storage center. There is no need to worry about the data loss problem caused by intrusion of virus and hacker and destruction of hardware by using Cloud Computing platform.

In the Internet era, the virus and hackers are rampant, data security and reliability is becoming more and more important. Contests which use cloud computing service, can store the data by the advanced data center in the world, have the strong technical management team to manage the submitted data, and provide reliable and safe data storage centers for the tournament. By using the cloud computing platforms, there is no need to worry about data loss problems resulting from viruses and attacks of hackers and hardware damage in the future contest areas

There are many issues that we have to explore because of the opening of cloud computing, high efficiency of equipment, extendibility, deployment flexibility of the business and so on.

ACKNOWLEDGEMENTS

This work is supported by National Nature Science Foundation (No 60873208) and Hebei Natural Science

Foundation (No F2009000927). The author would like to thank the reviewers of this paper for useful comments and suggestion.

REFERENCES

- [1] ACM International Collegiate Programming Contest.[EB/OL].(2009-10-16)[2010-09-10].<http://cm.baylor.edu/welcome.icpc/>.
- [2] UCSB International Capture The Flag.[EB/OL].(2007-12-18)[2010-10-05]. <http://ictf.cs.ucsb.edu/>.
- [3] Collegiate Cyber Defense Competition.[EB/OL].(2005-03-10)[2010-11-10].<http://www.nationalccdc.org/>
- [4] NTU Network Security Competition.[EB/OL].(2003-08-01)[2010-09-05].<http://www.lugs.org.sg/pipermail/slugnet/2003-August/006022.html>.
- [5] University NOC activities of network security competition. [EB/OL].(2009-06-01)[2010-11-15].<http://g.noc.net.cn/>.
- [6] J. Paulo Leal and Fernando Silva, "Mooshak: a Web-based multi-site programming contest system," Software: Practice and Experience, vol. 33, May 2003, pp. 567-581, doi: 10.1002/spe.522.
- [7] A.Shamsul Arefin, M. Arifur Rahman, S. Anwar Sharna, Samiran Mahmud, and Dr. M. Kaykobad, "Secured Programming Contest System with Online and Real-time Judgment Capability," 8th International Conference on Computer and Information Technology (ICIT) , Bangladesh, 2005, pp. 584-586.
- [8] A. Trotman and C. Handley, "Programming contest strategy," Computers Education, vol. 50, April 2008, pp. 821-837, doi:10.1016/j.compedu.2006.08.008.
- [9] R. Buyya, S. Pandey, and C. Vecchiola, "Cloudbus Toolkit for Market-Oriented Cloud Computing," Lecture Notes in Computer Science, vol 5931/2009, 2009, pp. 24-44, doi: 10.1007/978-3-642-10665-1_4.
- [10] Bhatiya Wickremasinghe, Rodrigo N. Calheiros , and R. Buyya, "CloudAnalyst: A CloudSim-based Visual Modeller for Analysing Cloud Computing Environments and Applications," 2010 24th IEEE International Conference on Advanced Information Networking and Applications, April 2010, pp. 20-23.
- [11] Rodrigo N. Calheiros1, Rajiv Ranjan1, César A. F. De Rose, and Rajkumar Buyyaetc. "CloudSim:A Novel Framework for Modeling and Simulation of Cloud Computing infrastructures and Services", echnical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory ,2009,3.
- [12] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges", Journal of Internet Services and Applications, vol. 1, Appl. 2010, pp. 7-18, doi: 10.1007/s13174-010-0007-6.
- [13] Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres, and Maik Lindner, "A Break in the Clouds: Towards a Cloud Definition", ACM SIGCOMM Computer Communication Review. New York, vol 39, pp. 50-55, January 2009.
- [14] Dragon-lab.[EB/OL].(2006-05-06)[2010-07-10].<http://www.Dragonlab.org/>.
- [15] JiLong Wang, ZhongHui LI, GuoHan LÚ, CaiPing Jiang, Xing Li, and QianLi Zhang, "DRAGON-Lab-Next generation internet technology experiment platform", Science in China Series F: Information Sciences, vol. 51, Nov. 2008, pp. 1908-1918, doi: 10.1007/s11432-008-0149-3.

Cloud Computing: Issues in Data Mobility and Security

Zaigham Mahmood

School of Computing and Mathematics
University of Derby
Derby, UK
e-mail: z.mahmood@derby.ac.uk

Harjinder Singh Lallie

School of Computing and Mathematics
University of Derby
Derby, UK
e-mail: h.s.lallie@derby.ac.uk

Abstract— Cloud Computing is a generic term for anything that involves delivering hosted services over the internet, based on a pay-as-you-go approach. Cloud Computing offers numerous benefits and, therefore, many large enterprises have embraced the cloud technologies and infrastructures. Vendors are also developing tools and applications to fulfill the demand and tap into the growing market. Like any new technology and paradigm, there are also numerous issues and concerns, including security and availability of data. This paper explores some of the security issues surrounding data location, mobility and availability as well as issues relating to the security of data at rest. The aim is to provide some useful background information for enterprises preparing to take advantage of Cloud Computing paradigm.

Keywords- cloud computing; enterprise computing; cloud security; data privacy; data security

I. INTRODUCTION

Cloud Computing is a generic term for anything that involves delivering hosted services and computing resources over the Internet. It is ‘a style of computing where massively scalable IT-enabled capabilities are delivered as services to external customers using Internet technologies’ [1]. According to NIST (National Institute of Standards and Technology, US), it provides ‘a convenient, on-demand network access to a shared pool of computing resources’ [2, 3]. Here, resources refer to computing applications, software services, platforms and computing infrastructures. Forrester [4] suggests that Cloud Computing refers to ‘a pool of abstracted, highly scalable and managed infrastructure capable of hosting end-customer applications and billed by consumption’. It is the latest paradigm in distributed computing that promises to revolutionize IT and business by making computing available as a *utility* over the World Wide Web.

General public have been using Cloud Computing in the form of Internet services like *Hotmail* (since about 1996), *YouTube* (since about 2005), *Facebook* (since about 2006) and *Gmail* (since about 2007). *Hotmail* is probably the first cloud computing application that allowed the general public to keep their data in the form of text and files at the vendor’s servers. Since then, many other services have emerged that allow users to store information (such as text files, photographs, video clips and music) and perform processing

without paying any upfront fees. Some well known examples include: *Twitter*, *Myspace*, *Wikipedia* and *Google docs*. These are typically consumer oriented services, different from enterprise-oriented tasks, but the underlying principles are the same i.e. to provide the storage space and processing capability. In the commercial arena, *Amazon.com* was one of the first vendors to provide storage space, computing resources and business functionality following the cloud computing model. *Salesforce.com*, founded in 1999, pioneered the concept of delivering enterprise applications as services to enterprises. In 2002, *Amazon Web Services* provided a suite of cloud-based services and, later in 2006, it launched *Elastic Compute Cloud* (EC2) that allowed companies and individuals to rent computers on which to run their own enterprise applications. The number of cloud services providers and the applications, platforms and infrastructures are increasing at such a rate that, in 2009, Gartner listed Cloud Computing as number 1 in its top 10 strategic technology areas for 2010 [10, 11]. The reasons why more and more companies are planning to take advantage of IT Cloud Computing solutions include:

- reduced costs associated with delivering IT services
- reduced management responsibilities
- increased business efficiency and agility
- easy access to software and hardware resources available elsewhere
- no long term contracts with vendors.

A report on Cloud Computing published in Jan 2010 [12] suggests that: 1) Enterprises are now moving beyond experimentation; 2) they are beginning to develop and deploy management software to deal with scaled Cloud environments; and 3) they are beginning to develop enterprise-level policies and standards for dealing with *Public* and *Hybrid* Clouds.

In the rest of this paper, we first outline the benefits that Cloud Computing offers and briefly discuss the deployment approaches. Then, in Sections III and IV, we discuss, in some detail, the inherent issues with respect to mobility and security of data held on Clouds. The last section presents a brief conclusion.

II. CLOUD COMPUTING

A. The Promise

Large vendors like IBM, Dell, Oracle and Sun have started to take strong positions with respect to Cloud Computing provision [5]. The essential features of this latest paradigm include [2, 6]:

- On-demand services: to enable users to avail of computing capabilities as and when required
- Resource pooling: to allow dynamically assigned computing resources to serve multiple consumers
- Rapid elasticity: to allow services, resources and infrastructures to be automatically provisioned
- Measured provision: to provide a metering capability to determine the usage for pricing purposes
- Effective management: to provide and facilitate easy monitoring, controlling and reporting.

Cloud Computing is an attractive paradigm that can be massively scalable. It provides benefits of efficiency, flexibility and high utilization that, in turn, can result in reduced capital investment costs and lower operational expenditure. The Cloud offerings from service providers and vendors are continuing to mature and increase in number. With this, the cost savings are becoming particularly attractive. There is no doubt that Cloud Computing is making *supercomputing* available to the masses.

B. Deployment Approaches

Cloud Computing can be classified and deployed in a number of ways e.g. as *public*, *private* or *hybrid* clouds.

Public Clouds are networks where services are provided by third parties and hosted and managed by the service providers. The Cloud providers take on the responsibilities of installation, management, provisioning and maintenance. Consumers are charged only for the resources they use following a pay-as-you-go model.

Private Clouds are proprietary networks normally residing within the enterprises for the exclusive use of the organization or for a known group of consumers. In case of Private Clouds, the enterprise is in charge of maintaining the Cloud and also responsible for security and regulatory compliance issues. The issues of data security are, therefore, somewhat reduced.

Hybrid Clouds are a combination of Private and Public Clouds. In this case, the management responsibilities are often split between the enterprise and the Public Cloud providers, which can often become an issue of concern. For mission critical processes, this type of Cloud infrastructure is much more effective because of enhanced control and management by the enterprise itself.

The Cloud model consists of, typically, three components which refer to three types of services: *Software Services*, *Platform Services* and *Infrastructure Services*. These services may be defined as follows:

- Software as a Service (SaaS): referring to prebuilt and vertically integrated applications available for purchase or use by customers as *services*. Here, customers are looking to 'hire' easy-to-consume functionality.
- Platform as a Service (PaaS): referring to application development toolkits and deployment tools (e.g. application servers, portal servers and middleware) which clients make use of to build and deploy their own applications. Here, customers are looking to buy time and cost savings in deploying applications.
- Infrastructure as a service (IaaS): referring to hardware (e.g. servers, storage space, network devices, etc) to enable Cloud Platforms and Applications to operate. Here, customers are looking to hire *computing*. Since, the infrastructure is offered on pay-for-what-you-use basis, it is sometimes referred to as *utility computing*.

C. Inherent Issues

Notwithstanding the benefits that Cloud Computing offers, there are numerous issues and challenges for organizations embracing this new paradigm. Zhen [8] lists a number of major challenges with respect to the following: 1) governance, management and updating of data; 2) management of software services; 3) monitoring of products and processes; 4) reliability and availability of systems and infrastructure and 5) security of information and data. The Expert Group Report [9] mentions a number of issues including: 1) concerns over security with respect to valuable knowledge, information and data placed on an external service; 2) concerns over availability and business continuity; and 3) concerns over data transmission across anticipated broadband speeds. Other shortcomings, as mentioned by various researchers, include: 1) no native security attributes; 2) inadequate or no security provisioning by providers; 3) lack of understanding of cloud legal issues; and 4) the failure to recognize potential liability from either legal issues or a lack of security. Issues with respect to "control" are also real concerns. A closer examination reveals that the major concerns may be broadly classified as those relating to the following:

- Security, including reliability and availability
- Governance and Management.

In this paper, we discuss issues with respect to data mobility, security and availability. Other issues are discussed in a companion papers which is under preparation.

III. DATA MOBILITY AND SECURITY

Cloud Computing provides services with respect to enterprise applications (software components and systems), computing platforms (development tools) and infrastructures (hardware including servers). Benefits are huge but the inherent issues are also many. Some of the major issues refer to the security of data. In this respect, there are

many dimensions including: data security, data privacy, data protection, data availability and data transmission. Forrester [22] combines these into three groups: 1) Security and Privacy; 2) Compliance; and 3) Legal and Contractual. Some of these are now discussed in some detail.

A. Data Mobility

Cloud Computing offers a high degree of data mobility in the sense that data stored on the Cloud may reside on a location geographically a long way away from the organization that owns the data. In a majority of cases, the owners and the users know where the data resides; however, this may not be true in all cases. Unless there is a contractual agreement that data should stay in a particular location or reside on a given known server, the Cloud providers may decide to keep it moving from one location to another. There are several reasons for this, including: 1) reducing the cost of storing data; 2) efficiency of retrieval of data; 3) easy availability of data; 4) efficient linking of different data resident on different locations; and 5) resource optimization. Security risks and issues are already big concerns. When data mobility is at a high level then the risks and issues increase many folds especially when data is transferred to another country with different regulations. High levels of data mobility also have negative implications for data security and data protection as well as data availability.

Many factors influence the choice of location for data centres including the cost as the cost of running a centre is high on the list of priorities [19]. The attraction of co-location and distribution of data is particularly justifiable due to the bandwidth efficiencies that this could provide. For instance Amazon's *CloudFront* data centres are located in the following cities: Ashburn Virginia, Dallas/Fort Worth, Los Angeles, Miami, Newark New Jersey, Palo Alto, California, Seattle, St. Louis, Amsterdam, Dublin, Frankfurt, London, Hong Kong and Tokyo. This does not necessarily mean that data stored on the Amazon Cloud may be split across any number of these data centres.

Another influencing factor is the cost relating to high capacity internet access.

B. Data Availability

Data availability is a major legitimate reason for the data to be stored in multiple locations on the Cloud. This answers a core business requirement: that of an uninterrupted service and seamless provision. Furthermore, data availability is such a crucial issue that it is common for Cloud providers to credit customer accounts if the system downtime duration drops below that specified in the SLA (service level agreement). The related issue is that, often, such measures are not specified in the SLAs.

The issue of data availability is exemplified by the outages suffered by Google's *Gmail* service in February 2009 which resulted in embarrassing headlines for the company [17]. In the subsequent service agreement for its

Premier Apps range of products which also covers *Gmail*, Subsequently, Google promised that customer data availability will be at least 99.9% of the time in any calendar month [18].

C. Cost Relating to Data Mobility

Another reason for data mobility is to reduce the cost of running data centres (by reducing the electricity bills, for example). In Public Clouds, data is often routed to other locations at certain times of the day or year, or when there is a huge climatic temperature fluctuation [20]. The main factor in such considerations is the cost of provision. Qureshi [21] has conducted research into the dynamic routing of data based on the cost of electricity in various regions. This research shows that it is possible to reduce electricity costs by up to 40%. However, as electricity costs rise, Cloud service providers may look for more effective ways of reducing their overheads – at the same time, hopefully, ensuring that there is no compromise on performance and service availability. In this respect, Qureshi's method [21] of dynamically routing data would become an attractive solution. From the point of view of data security and data availability, this would exacerbate the security issues, which are already a major concern when data is being moved between locations. As mentioned before, data mobility or dynamic data routing is also considered as a result of resource optimization. This, in turn, also helps to reduce costs.

Enterprises consuming Cloud services may not be aware of this, however, as they become more knowledgeable, they may decide to request appropriate data relocation and, thus, negotiate lower contract prices for such data services.

D. Data Location Assurance

Data mobility and location concerns, including those relating to security, have been partially addressed by Cloud providers, and two of the largest vendors in the field have started offering solutions to customers. Amazon's AWS (Amazon Web Services) provides an option within its *S3* (Simple Storage Service) package to allow customers to specify the regions for the storage and location of their data. It also provides assurance that data will not leave the customer selected regions [23]. Although, the available locations are currently restricted to just three regions: US (Standard), EU (Ireland) and US-West (Northern California), the company has plans to expand into the Asia-Pacific region in 2010. Amazon is marketing this as a way of improving performance and providing a better customer-centric service.

In 2009, Microsoft announced that its Windows *Azure* system would provide its users with the option to specify where in the world they would wish their data to be stored. As well as performance gains, Microsoft also stated legal and regulatory reasons for this facility. This is an attractive facility as different countries have different laws with respect to data privacy and confidentiality and some clients may wish to exploit such differences to their advantage, although

there may be legal implications. However, as with AWS, Microsoft has a very restricted choice of geographical locations, currently only two: both within the US. Microsoft has plans to expand this and is especially interested in sites outside the US.

E. Cross Border Data Transition

Cross border data transition can lead to potential legal risks due to different locations having varying policies, regulations and legislation. This means that data protected by legislation in one country may not have the same, or even similar, protection in another country [24]. In an example of this, that appears in Jaeger [19], it is noted that the European Union and United States of America have differing definitions of privacy as a result of disparate privacy policies. Their Data Protection Laws are based on the assumption that the location and responsibility of data is known and understood. Cloud Computing however challenges this presumption.

Presently, a vast majority of data centres are located in the United States [25]. A consequence of this is that data protection and privacy concerns are influenced by the USA Patriot Act 2001, the Foreign Intelligence Surveillance Act (FISA amendments act of 2008), the Electronic Communications Privacy Act (1986), the Privacy Act (1974) and the Homeland Security Act (2002). Under the ruling of these acts, the FBI and similar agencies have the regulatory power to demand access to any data stored on any computer within the USA, even if it is stored on behalf of another jurisdiction [25].

F. Organisations' Response

One of the key recommendations made by Gartner [26] suggests that assurance should be given guaranteeing that customer data will be stored and processed within a certain jurisdiction and that the local laws within that jurisdiction would apply. However, this may conflict with the concept of data privacy particularly in countries such as the US. This means that consumer data stored within the US may be highly vulnerable to disclosure [24] which may pose a potential business and/or economic risk.

World governments and organizations are developing strategies to counter these concerns e.g. the Canadian Government does not allow the use of hosting services that are based in the US [25]. Similarly, SWIFT, an international banking firm, are locating data centres in Switzerland where the data protection regulations are based on EU laws providing specific conditions for the transference of personal data to third parties or abroad [27].

The UK Data Protection Act 1998 requires that personal information is handled in a manner that ensures that key principles and legal obligations are properly adhered to. These principles also include the restriction on transferring data to countries outside of the European Economic Area (the EEA) unless there is a clear and adequate data protection mechanism [28]. The act also makes it an obligation for companies to clearly state where customers' data is being

held. However, this may be difficult when providers themselves are unaware of the exact locations.

IV. SECURING DATA AT REST

A valid question with respect to security of data on the Cloud is: how to ensure security of *data at rest*. The obvious answer suggests that data should be encrypted. Unfortunately, this is not as simple as it appears. If the data is being stored by an *IaaS* service (such as Amazon's *Simple Storage Service*, also known as S3), that is not associated with a specific application, then encryption is appropriate and indeed possible and a valid solution. However, data on the Cloud being processed by *SaaS* or *PaaS* applications (such as *Salesforce.com* or *Google Apps*) is generally not considered suitable for encryption. This is because encryption prevents indexing or searching of data, which has implications on availability and access of such data.

Finding an appropriate mechanism that is secure, and allows both addition and multiplication, has proved elusive and many respected cryptologists have suggested that it may not even be possible. However, a number of techniques and schemes have subsequently been put forward favoring the full homomorphic encryption [29]. A fully homomorphic cryptosystem is one where the performance of a mathematical operation on ciphertext is found to have a regular effect on the corresponding plaintext.

One such scheme, developed by Gentry [30], allows data to be processed without being decrypted. This means that a Cloud service provider can perform computations on client's data without exposing the original data. Gentry's proposal has caused great interest amongst cryptographers who have been trying to develop a practical manifestation of the concept of privacy homomorphism for over thirty years [16].

Other research efforts are focusing on methods to limit the amount of data that needs to be decrypted for processing in the Cloud. An example is predicate encryption, a type of asymmetric encryption where different individuals or groups can selectively decrypt some of the encrypted data instead of decrypting all of it [22]. Methodologies are being developed.

V. CONCLUSION

Cloud Computing is 'essentially on-demand access to a shared pool of computing resources'. It helps consumers to reduce costs, reduce management responsibilities and increase business agility. For this reason, it is becoming a popular paradigm and increasingly more companies are shifting toward IT Cloud Computing solutions. Advantages are many but there are also challenges and inherent issues. Generally, these relate to data governance, service management, process monitoring, infrastructure reliability, information security, data integrity and business continuity. This paper focuses on the mobility and availability of data held on the Cloud and discusses the security issues of such data.

In spite of the limitation and issues as discussed in the previous sections, Cloud Computing is becoming an attractive paradigm for large enterprises. In 2008, Forrester [5] predicted that ‘cloud computing initiatives could affect the enterprises within 2 to 3 years as it has the potential to significantly change IT’. In 2009, Gartner listed Cloud Computing as number 1 in its top 10 strategic technology areas for 2010 [10, 11]. In another report, Gartner suggested that ‘by 2012, 80% of Fortune companies will pay for some cloud computing service and 30% of them will pay for cloud computing infrastructure’ [4]. Enterprises are excited about the opportunities that Cloud Computing presents and, as the evidence suggests [4, 5, 10-12], Enterprise Cloud Computing is firmly poised to be the *next big thing* for businesses, large and small.

REFERENCES

- [1] David W Cearley, Cloud Computing: Key Initiative Overview, Gartner Report, 2010
- [2] Peter Mell and Tim Grance, The NIST Definition of Cloud Computing, version 15, National Institute of Standards and Technology (NIST), Information Technology Laboratory, www.csrc.nist.gov, 7 Oct 2009
- [3] Dustin Amrhein and Scott Quint, Cloud Computing for the Enterprise: Part 1: Capturing the Cloud, DeveloperWorks, IBM, 8 Apr 2009,
- [4] John Rhoton, Cloud Computing Explained: Implementation Handbook for Enterprises, Recursive Press, 3 May 2010
- [5] John M Willis, Cloud Computing and the Enterprise, IT Management and Cloud, [Online] Available at: www.johnmwillis.com/ibm/cloud-computing-and-the-enterprise/, 13 Feb 2008
- [6] Caroline Kvitka, Clouds Bring Agility to the Enterprise, [Online] Available at: http://www.oracle.com/technology/oramag/oracle/10-mar/o20interview.html
- [7] Michael Sheehan, Cloud Computing Expo: Introducing the Cloud Pyramid, Clod Computing Journal, Aug 2008
- [8] Jian Zhen, Five Key Challenges of Enterprise Cloud Computing, Cloud computing journal, 16 Nov 2008
- [9] Lutz Schubert, The Future of Cloud Computing, Expert Group Report, [Online] Available at: http://cordis.europa.eu/fp7/ict/ssai/docs/executivesummary-forweb_en.pdf
- [10] Dustin Amrhein & Scott Quint, Cloud Computing for the Enterprise: Part 1: Capturing the cloud, Understanding cloud computing and related technologies, DeveloperWorks, IBM, [Online] Available at: www.ibm.com/developerworks/websphere/techjournal/0904_amrheinn/0904_amrhein.html
- [11] Stephen Shankland, Brace yourself for Cloud Computing, CNET News, Oct 2009 http://news.cnet.com/8301-30685_3-10378782-264.html
- [12] Ravi Mhatre, Top 5 trends for enterprise cloud computing in 2010, Lightspeed Venuter Partners, Jan 2010
- [13] Sharon Sasson, Seven Best Practices for Cloud Computing, Enterprise Systems, August 2008, [Online] Available at: http://esj.com/articles/2009/08/18/cloud-best-practices.aspx
- [14] David Linthicum, Cloud Computing? Thank SOA, [Online] Available at: http://www.thecloudtutorial.com/cloud-computing-soa.html
- [15] David Linthicum, Cloud Computing and SOA Convergence in Your Enterprise: A Step-by-Step Guide, Addison Wesley, 2009
- [16] Katz J., Sahai, A., & Waters, B. (2008) Predicate Encryption Supporting Disjunctions, Polynomial Equations, and Inner Products, Proceedings of the theory and applications of cryptographic techniques 27th annual international conference on Advances in cryptology [Online] Available at http://eprint.iacr.org/2007/404.pdf. (Accessed 6th December 2009)
- [17] BBC, Google users hit by mail blackout, BBC News, 24 February 2009. [Online]. Available at http://news.bbc.co.uk/1/hi/technology/7907583.stm (Accessed: November 2009).
- [18] Google. Google Apps Service Level Agreement, 2009. [Online]. Available at: http://www.google.com/apps/intl/en/terms/sla.html (Accessed: November 2009).
- [19] Jaeger, P. T., Grimes, J. M., Lin, J. & Simmons, S, ‘Cloud Computing and Information Policy: Computing in a Policy Cloud?’ Journal of Information Technology & Politics, 5(3), 2008
- [20] Knight, W, Energy-Aware Internet Routing, 2009. [Online]. Available at: www.technologyreview.com/business/23248/page2/ (Accessed: November 2009)
- [21] Qureshi, A, Plugging Into Energy, 7th ACM Workshop on Hot Topics in Networks (HotNets). Calgary, Canada, October 2008
- [22] Wang C, Cloud Security Front and Centre, Forrester Report, Nov 2009
- [23] Amazon Web Services, Amazon Simple Storage Service FAQs, 2009. [Online]. Available at: http://aws.amazon.com/s3/faqs/#Where_is_my_data_stored (Accessed: 9 December 2009)
- [24] European Network and Information Security Agency, (2009) Cloud Computing, Benefits Risks and Recommendations for Information Security, [Online] Available at: http://enisa.europa.eu/
- [25] Thompson, B, Storm warning for cloud computing, BBC News, 17 May 2008 [Online]. Available at: http://news.bbc.co.uk/1/hi/7421099.stm (Accessed: November 2009)
- [26] Gartner, Assessing the Security Risks of Cloud Computing, 2008,[Online] Available at: http://www.gartner.com/DisplayDocument?id=685308 Last accessed: 7th December 2009
- [27] Economist, Computers without borders, 2008, [Online] Available at: Economist (23 October), at http://www.economist.com/ , Last accessed: 6th December 2009
- [28] ICO, Review of EU Data Protection Directive: Summary, 2009, [Online]. Available at: http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/review_of_eu_dp_directive_summary.pdf Date of access: (28 November 2009)
- [29] Benaloh J., Verifiable Secret-Ballot Elections. PhD thesis, Yale University, 1987.

- [30] Gentry, C, Fully homomorphic encryption using ideal lattices, Annual ACM symposium on theory of computing, Proceedings of the 41st annual ACM symposium on theory of computing, Bethesda, MD, USA, 2009, Session: crypto, pp 169-178, ACM, New York, NY, USA.
- [31] Rivest, R., Adleman, L., & Dertouzos, M., On data banks and privacy homomorphisms. In Foundations of Secure Computation, pp. 169–180, 1978.