ICIW 2012

The Seventh International Conference on Internet and Web Applications and Services

ISBN: 978-1-61208-200-4

May 27- June 1, 2012

Stuttgart, Germany

ICIW 2012 Editors

Friedrich Laux, Reutlingen-University, Germany

Pascal Lorenz, University of Haute Alsace, Colmar, France

ICIW 2012

Forward

The Seventh International Conference on Internet and Web Applications and Services (ICIW 2012) held on May 27 - June 1, 2012 - Stuttgart, Germany, 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 2012 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.

We take this opportunity to thank all the members of the ICIW 2012 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 2012. 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 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICIW 2012 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 city of Stuttgart surely provided a pleasant environment during the conference and we hope you had a chance to visit the surroundings.

ICIW 2012 Chairs

Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Mihhail Matskin, NTNU, Norway
Vagan Terziyan, University of Jyvaskyla, Finland

ICIW 2012

Committee

ICIW Advisory Committee

Steffen Fries, Siemens AG, Corporate Technology - Munich, Germany
Mihhail Matskin, NTNU, Norway
Vagan Terziyan, University of Jyvaskyla, Finland

ICIW 2012 Technical Program Committee

Mehmet Aktas, Indiana University, USA
Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Markus Aleksy, ABB Corporate Research Center, Germany
Giner Alor Hernandez, Instituto Tecnologico de Orizaba - Veracruz, México
Eckhard Ammann, Reutlingen University, Germany
Cristobal Arellano, University of the Basque Country UPV/EHU, Spain
Khedija Arour, University of Carthage - Tunis & El Manar University, Tunisia
Marzieh Asgarnezhad, Islamic Azad University of Kashan, Iran
Nahed A. Azab, The American University in Cairo, Egypt
Ana Sasa Bastinos, University of Ljubljana, Slovenija
Siegfried Benkner , University of Vienna, Austria
Giancarlo Bo, Technology and Innovation Consultant- Genova, Italy
Christos Bouras, University of Patras / Research Academic Computer Technology Institute, Greece
Laure Bourgois, INRETS, France
Mahmoud Brahimi, University of Msila, Algeria
Ruth Breu, University of Innsbruck, Austria
Mihaela Brut, IRIT, France
Dung Cao, Tan Tao University - Long An, Vietnam
Miriam A. M. Capretz, The University of Western Ontario - London, Canada
Ajay Chakravarthy, University of Southampton, UK
Xi Chen, Nanjing University, China
Dickson Chiu, Dickson Computer Systems, Hong Kong
María Consuelo Franky, Pontificia Universidad Javeriana - Bogotá, Columbia
Javier Cubo, University of Malaga, Spain
Roberta Cuel, University of Trento, Italy
Richard Cyganiak, Digital Enterprise Research Institute / NUI Galway, Ireland
Paulo da Fonseca Pinto, Universidade Nova de Lisboa, Portugal
Maria Del Pilar illamil Giraldo, Universidad de los Andes, Columbia
Gregorio Diaz Descalzo, University of Castilla - La Mancha, Spain
Matei Dobrescu, Insurance Supervisory Commission, Romania
Eugeni Dodonov, Intel Corporation- Brazil, Brazil
Ioan Dzitac, Aurel Vlaicu University of Arad, Romania
Matthias Ehmann, University of Bayreuth, Germany

Javier Fabra, University of Zaragoza, Spain
Jacques Fayolle, Télécom Saint-Etienne/l'Université Jean Monnet, France
Adrián Fernández Martínez, Universitat Politecnica de Valencia, Spain
Chiara Francalanci, Politecnico di Milano, Italy
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, IDRBT, India
Mouzhi Ge, Bundeswehr University Munich, Germany
Jean-Pierre Gerval, ISEN Brest, France
Mohamed Gharzouli, Mentouri University of Constantine, Algeria
Katja Gilly, Universidad Miguel Hernández, Elche, Alicante, Spain
Gustavo González-Sanchez, Mediapro Research, Spain
Feliz Gouveia, Universidade Fernando Pessoa - Porto, Portugal
Anna Goy, Università di Torino, Italy
Andrina Granić, University of Split, Croatia
Sven Graupner, Hewlett-Packard Laboratories - Palo Alto, USA
Carmine Gravino, Università degli Studi di Salerno, Italy
Patrizia Grifoni, CNR-IRPPS, Italy
Bidyut Gupta, Southern Illinois University - Carbondale, USA
Ileana Hamburg, Institut Arbeit und Technik, Germany
Nael Hirzallah, Fahad Bin Sultan National University, Kingdom of Saudi Arabia
Chi Chi Hung, Tsinghua University - Beijing, China
Edward Hung, Honk Kong Polytechnic University, Hong Kong
Rauf Irum, Åbo Akademi University, Finland
Linda Jackson, Michigan State University, USA
Ivan Jelinek, Czech Technical University, Czech Republic
Monika Kaczmarek, Poznan University of Economics, Poland
Hermann Kaindl, Vienna University of Technology, Austria
Jalal Karam, Alfaisal University-Riyadh, Kingdom of Saudi Arabia
Brigitte Kerherve, UQAM, Canada
Suhyun Kim, Korea Institute of Science and Technology (KIST), Korea
Alexander Knapp, Ludwig- Maximilians-Universität München, Germany
Kenji Kono, Keio University, Japan
Longzhuang Li, Texas A&M University-Corpus Christi, USA
Shiguo Lian, Orange Labs Beijing, China
Malamati Louta, University of Western Macedonia - Kozani, Greece
Zaigham Mahmood, University of Derby, UK
Zoubir Mammeri, IRIT - Toulouse, France
Chengying Mao, Jiangxi University of Finance and Economics, China
Mihhail Matskin, NTNU, Norway
Inmaculada Medina-Bulo, Universidad de Cádiz, Spain
Stephanie Meerkamm, University of Bayreuth, Germany

Andre Miede, BearingPoint/Management & Technology Consultants - Hamburg, Germany
Fernando Miguel Carvalho, Lisbon Superior Engineering Institute, Portugal
Serge Miranda, University of Nice, France
Sanjay Misra, Federal University of Technology - Minna, Nigeria
Nader Mohamed, UAE University, United Arab Emirates
Shahab Mokarizadeh, Royal Institute of Technology (KTH), Sweden
T.R. Gopalakrishnan Nair, Prince Mohammad Bin Fahd University, KSA
Alex Ng, The University of Ballarat, Australia
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, Tilburg University, The Netherlands
Federica Paganelli, CNIT - National Consortium for Telecommunications - Firenze, Italy
Helen Paik, University of New South Wales, Australia
Matteo Palmonari, University of Milan - Bicocca, Milan, Italy
Apostolos Papageorgiou, Technische Universitaet Darmstadt, Germany
João Paulo Sousa, Instituto Politécnico de Bragança, Portugal
Fredrik Paulsson, Umeå University, Sweden
George Pentafronimos, University of Piraeus, Greece
Mark Perry, Faculty of Law /Faculty of Science, University of Western Ontario - London, Canada
Stefan Pietschmann, Technische Universität Dresden, Germany
Marc Pous Marin, Barcelona Digital Center Tecnologic, Spain
Muthu Ramachandran, Leeds Metropolitan University, UK
Lucia Rapanotti, The Open University - Milton Keynes, UK
José Raúl Romero, Universidad de Córdoba/Campus de Rabanales, Spain
Christoph Reinke, SICK AG, Germany
Werner Retschitzegger, University of Linz, Austria
Jan Richling, Technical University Berlin, Germany
Gustavo Rossi, Universidad Nacional de La Plata, Argentina
Antonio Ruiz Martínez, University of Murcia, Spain
Fatiha Sadat, Université du Québec à Montréal, Canada
Saqib Saeed, University of Siegen, Germany
Muhammad Mohsin Saleemi, Åbo Akademi University, Finland
Sébastien Salva, IUT d'Aubière, France
David Sánchez Rodríguez, University of Las Palmas de Gran Canaria (ULPGC), Spain
Brahmananda Sapkota, University of Twente, The Netherlands
Antonio Sarasa-Cabezuelo, Complutense University of Madrid, Spain
Stefan Schulte, Vienna University of Technology, Austria
Wieland Schwinger, Johannes Kepler University Linz, Austria
Didier Sebastien, ESIROI-STIM, Reunion Island
Véronique Sebastien, University of Reunion Island, Reunion Island
Omair Shafiq, University of Calgary, Canada
Jawwad Shamsi, National University of Computer & Emerging Sciences - Karachi, Pakistan
Jun Shen, University of Wollongong, Australia
Patrick Siarry, Université Paris 12 (LiSSi) - Créteil, France

José Luis Sierra-Rodriguez, Complutense University of Madrid, Spain
André Luis Silva do Santos, Insituto Federal de Educação Ciencia e Tecnologia do Maranhão-IFMA, Brazil
Florian Skopik, AIT Austrian Institute of Technology, Austria
Vladimir Stancev, SRH University Berlin, Germany
Luis Javier Suarez Meza, University of Cauca, Colombia
Yuqing Sun, Shandong University, China
Jani Suomalainen, VTT Technical Research Centre of Finland, Finland
Sayed Gholam Hassan Tabatabaei, Universiti Teknologi Malaysia (UTM) - Kuala Lumpur, Malaysia
Panagiotis Takis Metaxas, Wellesley College, USA
Nazif Cihan Tas, Siemens Corporate Research - Princeton, USA
Vagan Terziyan, University of Jyvaskyla, Finland
Pierre Tiako, Langston University - Oklahoma, USA
Leonardo Tininini, ISTAT-Italian Institute of Statistics), Italy
Massimo Tisi, INRIA & École des Mines de Nantes, France
Konstantin Todorov, Ecole Centrale Paris, France
Giovanni Toffetti Carughi, University of Lugano, Switzerland
Orazio Tomarchio, University of Catania, Italy
Victor Manuel Toro Cordoba, University of Los Andes - Bogotá, Colombia
Samyr Vale, Federal University of Maranhão - UFMA - Brazil
Dirk van der Linden, Artesis University College of Antwerp, Belgium
Iván P. Vélez-Ramírez, Phidelix Technologies, Puerto Rico
Michael von Riegen, University of Hamburg, Germany
Rusen Yamacli, Anadolu University, Turkey
Beytullah Yildiz, Tobb Economics and Technology University, Turkey
Amelia Zafra, University of Cordoba, Spain
Weiliang Zhao, Macquarie University, Australia
Martin Zimmermann, Hochschule Offenburg - Gengenbach, Germany
Christian Zirpins, Karlsruhe Institute of Technology, Germany

Copyright Information

Table of Contents

A Validation Framework for the Service-Oriented Process Designing

Guoqiang Li[1,2], Lejian Liao[1], Fuzhen Sun[1]
[1]Beijing Engineering Research Centre of High Volume Language Information Processing & Cloud Computing Applications,
Beijing Key Laboratory of Intelligent Information Technology,
School of Computer Science, Beijing Institute of Technology, Beijing, China
[2]School of information, Linyi University, Linyi, China
{lgqsj, liaolj, 1090723}@bit.edu.cn

***Abstract*—In the service-oriented software systems, the services composition process is modeled using the service orchestration languages whose fault-handling and compensation mechanisms are crucial to guarantee the process running successfully. In this paper we propose to extend the syntax of BPEL to improve these two mechanisms. In order to validate their correctness, the composition is transformed to the planning graph. Then the validation of the fault-handling mechanism is regarded as a problem of seeking solution from the solution sets gained from the planning graph. We analyze the services composition structures and construct a relationship matrix to complete the validation of the compensation mechanism. A validation framework is proposed and an experiment is implemented to show our method effectiveness.**

Keywords - graph planning; BPEL; service orchestration; fault handling; compensation mechanism

I. INTRODUCTION

Service-oriented paradigm is capturing a growing interest as a mean for business to business integration. To realize the composition of the web services, researchers and industrial practitioners have proposed several web service orchestration languages such as BPEL4WS [1], WSFL [2], XLANG [3] and StAC [4]. And BPEL4WS is the de facto standard. Compare to the other languages, BPEL4WS supports this problem with a programmable and scope-based fault-handling and compensation mechanisms. Fault-handling mechanism guarantees the composition continues to achieve the goal. The function of compensation mechanism is to maintain the consistency of the whole process by eliminates the effects of everything executed from the failed service. But, it is a time-consuming and error-prone task to design these strategies and it is difficult to validate the correctness by the designers completely.

In order to solve the above problem, we propose to make use of graph planning technology focusing on the correctness validation of the business process during the design phase. The contribution of this paper includes:

- Syntax of the BPEL is proposed to extend with two operators corresponding to the fault-handling and compensation-handling mechanisms and their semantics are presented. A planning graph is constructed by means of analyzing the business process.
- For the fault-handling mechanism, it is transformed to seek a solution from the solution set of graph.
- Analyze the structural relationship of services and build a relationship matrix to facilitate the validation of the compensation-handling mechanism.

This paper is organized as follows. In the next section, we place the related work. Section Ⅲ introduces the extension of BPEL and the graph planning technology. The Section Ⅳ details the validation framework including the validate algorithms. The experiment is implemented and analyzed in Section Ⅴ. The conclusion of this paper and future work are discussed in the Section Ⅵ.

II. RELATED WORK

To guarantee the correctness of the business process, many researchers consider the semantic model. A simplified version of the WS-BPEL is defined in [5]. Compensation closure and context are proposed to capture the execution structure and form a good framework to the semantics of implementation of BPEL4WS. In [6], Chenguang, Shengchao and Zongyan verify the process using the Hoare-logic. In [7], Huibiao, Jifeng, Jing and Bowen focus on deriving the operational semantics and denotational semantics from algebraic semantics. Algebraic laws for BPEL programs are considered. Comparing with these methods, our approach is more intuitive.

A logic model specifies the semantics of workflows and composite tasks are given in [8]. A set of inference rules are presented to deduce the strongest post condition and weakest precondition and automatic workflow verification is demonstrated. The interactions of composite web services are modeled as conversations in [9]. The guarded automaton augmented with unbounded queues for incoming messages is used to be the intermediate representation and the model checker SPIN verifies synchronous communication. But, it is a challenge to translate the BPEL to the Promela program which is the input of SPIN for the designers. A Petri-net based formalization to construct composition process is proposed in [10]. And the interface dependency, compensation dependency and sequence triggered in nesting scopes are discussed. These preceding methods focus on the validation of the fault-handling and compensation-handling during the running phrase. From a transactional perspective of the compositions, many works introduce their approaches [11-13], e.g., a heuristic-based analysis of the process definition is proposed in [11]. The analysis result is a set of nonrepairable activities, whose impacts are evaluated by a repairability reasoner. Then a combination of the fault and the branching probabilities associated with an activity is given to gain a relevance index, which is used to remind the designer of knowing that to improve the repairability of the process. A transactional service patterns are used in [12] to specify the transactional composite service (CS) using Event Calculus. The CS transactional behavior is specified initially by the designers. Then the patterns and transactional flow are rewritten using EC predicates. Last, the behavior consistency is checked according to the predefined transactional consistency rules. Similar to our work, a planning graph is also used in [14], which only considers the repair technique for the composition adaptation rather than validating the correctness of the composition. A testing tool for web services composition is proposed in [15]. This tool focuses on conformance testing and unit testing considering the timing constraints and synchronous time delay. But, the activities of a flow activity are processed as sequence activities instead of processing in parallel in this tool. A web services translation tool is proposed in [16]. This tool is used to design and verify a web services system with time restrictions during the design phase. UML is used in the design phase to model the system to provide sequence diagram, which is transformed to choreography description by WS-CDL. Last, the UPPAAL tool is used for validation and verification purpose. An on-line approach is introduced in [17] to test an orchestration of web service composition and a passive testing verifies a timed trace with respect to a set of constraints. But, it does not pay close attention to the fault-handling and compensation-handling mechanisms.

III. EXTENSION OF BPEL AND GRAPH PLANNING

A. Syntax and Semantic of the Extended BPEL(ex-BPEL for short)

ex-BPEL builds on the base of the BPEL by extending the original fault handling mechanism. A business process (BP) includes four components: an activity P, a basic activity A, a fault handler F and a compensation handler C. The detailed syntax is as follows:

```
BP:= 【P: F】
P:= A        (basic activities)
 | skip          (do nothing)
 |P; P           ( sequence)
 |P || P         (flow)
 | if b then P else P   (conditional)
 |n: {P?C:F}       (scope)
A:= e           (assignment)
 | rec p y        (receive)
 | inv p x y        (invoke)
 | rep p x         (reply)
 | throw          ( throw a fault)
C, F := ↤n   ( compensation) | retry P: N | substitute P: P’
```

The operational semantics of P and A are same as the semantics of [6]. For example, sequence “P_1; P_2” presents an order of these two activities, i.e., P_2 starts running only after P_1 completes.

The extension of the fault mechanism includes two new operators: *retry* and *substitute*. The operator *retry* $P:N$ means activity P makes N repetitions and *substitute* $P_1:P_2$ means P_1 substitutes P_2 if P_2 fails. Actually, the two operators can be combined to describe complex handling strategy.

A work-through scenario is an e-travel example. To plan to travel from place A to B, a train ticket should be ordered first and another choice is to book a flight ticket if no train ticket. Then a hotel should be booked. In case of hotel booking failure, we can re-order the ticket or cancel the plan.

More details of the fault handling and compensation handling syntax of BPEL is referred to [1].

B. Extension of Fault-handling Mechanism

We distinguish two types of faults: temporary faults and permanent faults. For example, a temporary fault may be a network interruption in a short time. After the fault is thrown from the business process, firstly we analyze the type of the fault, and then we choose the handling mechanism for it. *Retry* is used to cope with the temporary faults, and *substitute* handles the permanent faults. So, the modified fault handler is as follows:

```
<faultHandlers>
<catch faultName="FailofTrainTicket"?
faultVariable="ncname"? >
<retry><invoke partnerLink=”TrainSupplier”
portType=”Trainsup:OrderInterface”
Operation=”submitOrder” inputVariable = “OrderInfo”
outputVariable= “OrderConfirmation”>[N]
</retry>
<substitute> <invoke partnerLink=”FlySupplier”
portType=“Flysup:OrderInterface”
Operation=”submitOrder”  inputVariable = “OrderInfo”
outputVariable= “OrderConfirmation”>
</invoke> </substitute></catch> </faultHandlers>
```

“[N]” specifies the number of repetitions in *retry* operation.

C. Introduction to Graph Planning

A planning graph is a directed, leveled graph with nodes and edges, denoting as $\langle V, E \rangle$ [18]. $V = \langle Prop, Action \rangle$, $Prop$ is a set of all proposition levels $\{\text{Prop}_0, \text{Prop}_1, \text{Prop}_2 ... \text{Prop}_n\}$, *Action* is a set of all action levels $\{\text{Action}_0, \text{Action}_1, \text{Action}_2 ... \text{Action}_n\}$ where an action is described as: $Action = (name(Params, Pre, Add, Del))$, where Pre specifies the preconditions of this action and Add specifies its positive effects. While the Del specifies its negative effects. The proposition levels and the action levels occur alternately. So, the planning graph is: $\{\text{Prop}_0, \text{Action}_0, \text{Prop}_1, \text{Action}_1 ... \text{Prop}_n\}$ shown in Fig. 1, where Prop_0 specifies the initial proposition level and Prop_n specifies the goals proposition level. If one proposition Prop0i exists in Pre of one action A, then there is an edge between Prop0i and A. Similarly, if one proposition Prop0j exists in Add of one action B, then there is an edge between Prop0j and B.

In Fig. 1, the black circle is a proposition node and the rectangle is an action node. The dotted line means every proposition that appears in proposition-level i may also appear in proposition level i+1, allowed by "no-op actions". Because of this trait, action-level i may contain all the possible actions whose preconditions all exist in proposition-level i [18]. So, for the *retry* operation, we can determine the maximum occurs times according to the N of a service when it fails. As shown in Fig. 1, the grey rectangle means the services ws_1 and ws_i should be updated in that level. We will not distinguish the action from service from here.

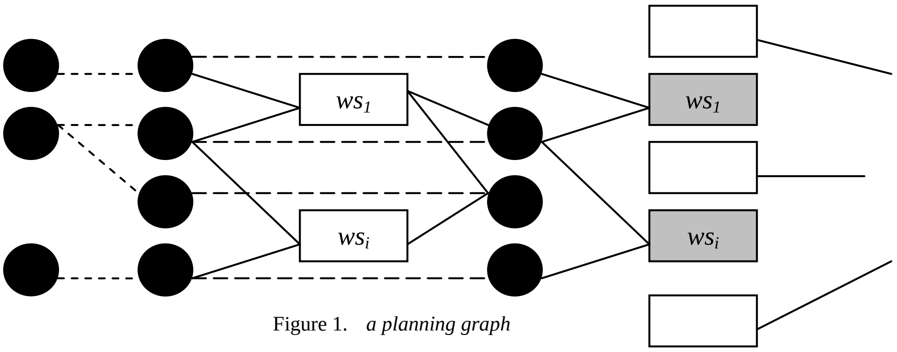


Figure 1. *a planning graph*

IV. VALIDATION FRAMEWORK

There are three modules in our validation framework as shown in Fig. 2. The Parsing module includes BPEL Parser and WSDL Parser. The former parses the BPEL documents and gets the service structure relationship matrix which is stored in the database. The latter parses the WSDL documents to get the corresponding actions. The Graph Planner is used to gain the solutions.

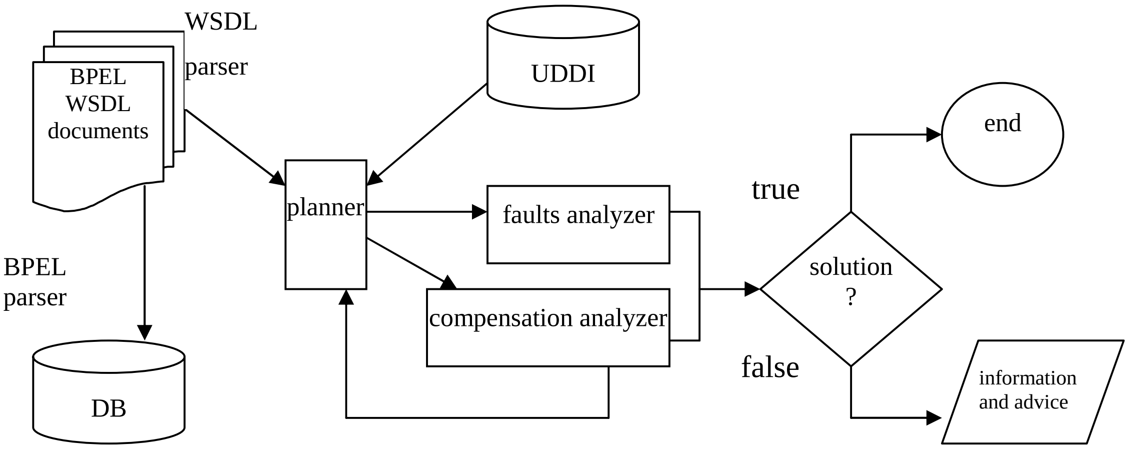


Figure 2. the validation framework

A. *Parsing Modules*

The parsing modules are responsible for generating the original input data.

1) *WSDL Parser*

WSDL is an XML format for describing web services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information. WSDL Parser transforms the services description to the actions presented by STRIPS, the algorithm is as follows:

```
input: WSDL documents:{$Wsdl_1$,$Wsdl_2$……}
output: a set of actions of planning: {$Action_1$,$Action_2$……}
procedure:
(a) name = the names appear in label: <wsdl:service></wsdl:service> of $Wsdl_i$
(b) Params = the list of all the names of the labels: <wsdl:message></wsdl:message>of $Wsdl_i$.
(c) Pre = the conjunction of all <wsdl:input></wsdl:input> as defined in label<wsdl:binding></wsdl:binding> of $Wsdl_i$.
(d)Add = the conjunction of all <wsdl:output></wsdl:output> as defined in label <wsdl:binding></wsdl:binding> of $Wsdl_{i.}$
(e) return $Action_i$ = (name(Params, Pre, Add)).
```

2) *BPEL Parser*

There are two ways in which two actions are marked to be exclusive of each other: (1) interference: if either of the actions deletes a precondition or add-effect of the other; (2) competing needs: if there is a precondition of action A and a precondition of action B that are marked as mutually exclusive of each other in the previous proposition level [18].

We treat the compensation information as the *del* information corresponding to the attribute "*del*" of the original *STRIPS* representation. Suppose *ws* is a service in business process, the set of services in its fault handler is defined $Wsf = \{wsf_1, wsf_2 ...\}$ and the set of services in its compensation handler is $Wsc = \{wsc_1, wsc_2 ....\}$. So, its mutual exclusion set is $Mes = Wsc$. All this information is stored in database and used in graph planning.

B. *Implements of Validation*

1) Construction of Planning Graph

It is simple to transform the business process to a planning graph. The actions of every level correspond to the services of one step of the process.

2) Definitions of Validation Properties

The validations of the fault and compensation handling mechanisms require all the satisfied solution i.e., the actions sequence: $S = \{S_1, S_2, S_3 ...\}$. We give some definitions on the validation properties.

a) fault-amendable service: for every service *ws*, if ws' exists and satisfies: $ws' \in Wsf$ and $ws' \in S$, we say this service fault-amendable.

b) fault-amendable process: if all the services of a process are fault-amendable, we say the process is fault-amendable.

c) compensation-amendable service: for every service *ws*, if ws' exists and $ws' \in Wsc$ can bring the process to a consistent state, we say this service compensation-amendable.

d) compensation-amendable process: if all the services of a process are compensation-amendable, we say the process is compensation-amendable.

e) reliable process: if the process is fault-amendable and compensation-amendable, we say the process is reliable.

3) Validation Algorithm of Fault-handling

```
begin
for each $ws_i \in BPEL$ Process
  if $\exists$ ws $\in Wsf \wedge ws \in S$
       return true
   else return $ws_i$
end
```

If the definition (a) is satisfied, the result is true, else a fault service is returned, and a handling suggestion will be given to the designers.

4) Validation of the Compensation-handling

According to the structures of the BPEL, we define the structure relationship as follows:

a) Sequence structure: in Fig. 3, ws_1 is a directly prior of ws_2, denoted: $ws_1 \prec ws_2$. And ws_2 is a directly successor of ws_1, denoted: $ws_2 \succ ws_1$. If a compensation of ws_2 is invoked, the compensation of ws_1 should be invoked.

b) Xor structure: in Fig. 4, ws_1 is a directly xor-split prior of ws_2 denoted: $ws_1 \prec_{xs} ws_2$. And ws_2 is a directly *xor-split* successor of ws_1, denoted: $ws_2 \succ_{xs} ws_1$. ws_2 is a directly *xor-join* prior of ws_4, denoted : $ws_2 \prec_{xj} ws_4$, ws_4 is a directly *xor-join* successor of ws_2, denoted: $ws_4 \succ_{xj} ws_2$.

c) And structure: in Fig. 5, ws_1 is a directly *and-split* prior of ws_2 denoted: $ws_1 \prec_{as} ws_2$ .And ws_2 is a directly *xor-split* successor of ws_1, denoted: $ws_2 \succ_{as} ws_1$. ws_2 is a directly *and-join* prior of ws_4, denoted: $ws_2 \prec_{aj} ws_4$, ws_4 is a directly *and-join* successor of ws_2 ,denoted $ws_4 \succ_{aj} ws_2$.

d) Parallel-or structure: in Fig. 4, ws_2 and ws_3 are parallel in *xor* structure, denoted $ws_2 \|_{or} ws_3$. The pair of services will not affect each other while any of them throws a fault. In this case, if the compensation of ws_2 is invoked and ws$_3$ runs normally, the compensation of ws_1 can not be invoked.

e) Parallel-and structure: in Fig. 5, ws_2 and ws_3 are parallel in *and* structure, denoted: $ws_2 \|_{and} ws_3$. If ws_2 is compensated, ws_3 must be compensated, and the compensation of ws_1 will be invoked.

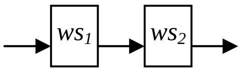


Figure 3. *Sequence structure*

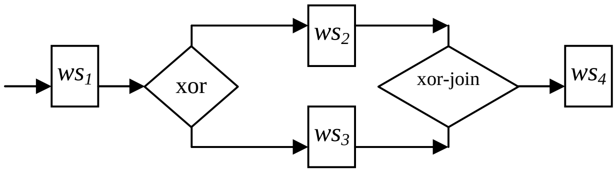


Figure 4. *Xor structure*

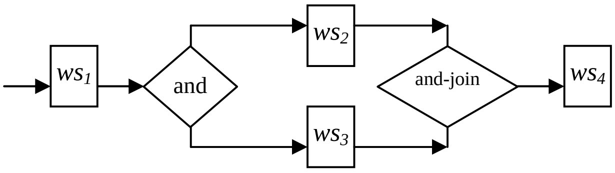


Figure 5. "*and*" *structure*

$$\begin{array}{ccccc} & ws_1 & ws_2 & ws_3 & ws_4 \\ ws_1 & - & \prec_{xs} & \prec_{xs} & - \\ ws_2 & \succ_{xs} & - & \|_{or} & \prec_{xj} \\ ws_3 & \succ_{xs} & \|_{or} & - & \prec_{xj} \\ ws_4 & - & \succ_{xj} & \succ_{xj} & - \end{array}$$

Figure 6. *Relationship matrix*

All the structure relationship of the services will be stored in a matrix, which can be automatically generated in the parsing of BPEL documents. The corresponding matrix of Fig. 4 is the Fig. 6, where the symbol "-" means no relationship between services.

Suppose the compensated service is *ws* and the service relationship matrix is *ws*-matrix. The algorithm to validate the compensation-handling mechanism is as follows:

a) Step 1: Look for a set of the services which is related to *ws* until a *xor-split* service *ws-xs* or the first service is found, and the path is recorded, denoted *ws-path*.

b) Step 2: Locate the directly successor *ws-post* of *ws*.

c) Step 3: Make sure whether there is a path from *ws-post* to *ws*-xs and the path exists in solution sets, if so, the validation of compensation-handling ends.

d) Step 4: if not, take the *del* information of the compensation services on the *ws-path* as the goal propositions and do the planning. If the solution can be found, the process is reliable, else the faulty service is located and advice is given.

For example, if ws_2 is compensated, its directly successor is ws_4. Because there are two path from ws_4 to ws_1, i.e., $ws_4 \rightarrow ws_2 \rightarrow ws_1$ and $ws_4 \rightarrow ws_3 \rightarrow ws_1$. So, if the compensation handling of ws_2 is defective nevertheless ws_3 is available, the process can run successfully.

5) Analysis of the Algorithms:

a) For the faults-handling, the complexity is O(n), n is the number of the services which is semantic or functionally equivalent to the faulty service in the same level.

b) For the compensation-handling, the complexity is O(n^2), which is the time needed to look for a path between given two nodes in a graph. If the path does not exist, we should do a new planning process, which is at least PSPACE-hard. In spite of this, the planning graph analysis can provide a quite substantial improvement in running time [18].

V. EXPERIMENT

The goals of the experiments are: (1) To validate the soundness and completeness. Soundness is that if there is a problem in the fault-handling and compensation-handling, the system is able to find it. Completeness is that if the fault information is returned, it is related to the designing. (2) To validate the efficiency of the algorithms. Because in our framework, wsdl4j is used to parse the WSDL documents, and dom4j is used to parse BPEL documents. So, we now focus on the validation efficiency of our proposed algorithms. The validation program is completed with Java on the platform Eclipse.

We adopt the dataset from [14]. There are 351 available services which use 2891 parameters in their input and output messages. This dataset has four groups, where Group 1 and Group 2 are chosen in our experiment. Group 1 contains solutions with 9 levels and Group 2 contains solutions with 18 levels. In our experiments, a random service of every level is presumed to be failed. At each experiment, we run the validation algorithms and running time is recorded. At Last, each data point is obtained from the average of three runs for the different failed service.

For the validation of fault-handling shown in Fig. 7, we change the size of the set *Wsf* of the failed service. Overall, the maximum running time is less than three milliseconds even though we set the size 100. Comparing the Group 1 with Group 2, there is not quite a difference in the running time in spite the fact that the levels of Group 2 is twice as many as the levels of Group 1. The main source of this conclusion is that the running time does not depend on the level of the service but the size of the set *Wsf* of the failed service. For designers, the running efficiency is quite acceptable.

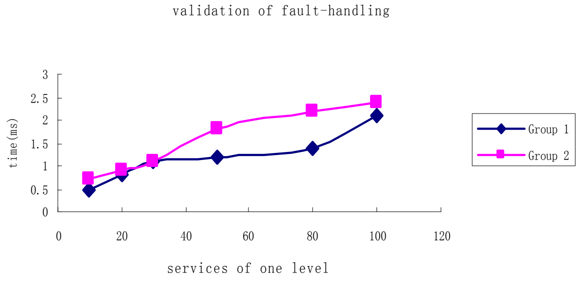
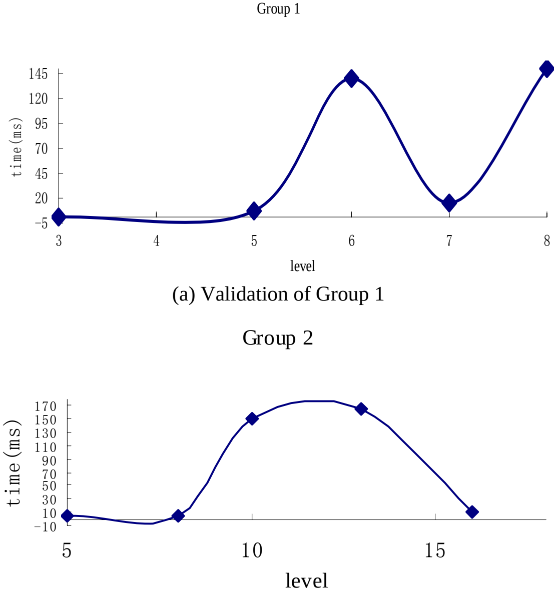


Figure 7. Validation of fault-handling



(a) Validation of Group 1

(b) Validation of Group 2

Figure 8. Validation of compensation-handling

For the validation of compensation-handling, because of their different levels, we place the running results of Group 1 and Group 2 into two figures, i.e., Fig. 8.a and Fig. 8.b. respectively. In the case of several paths existence from the faulty service to a consistent service, the running efficiency is very excellent. For example, the running time is only about 10 milliseconds even though the faulty service is in the tenth layer in the Fig. 8.b. But, it is very time-consuming to find a solution according to the del

ISBN: 978-1-61208-200-4

information of the faulty service. For example, the running time reaches 164 milliseconds in the thirteenth layer. At the same time, we can observe that it will take more time to make the validation when the faulty service is in the later layer under the same conditions. For example, the time taken in the thirteenth layer is longer than the tenth layer.

From the above analysis, it is feasible to take use of our method and it is acceptable for the designers.

VI. CONCLUSION

In this paper, we focused on the validation of the correctness of the service composition process during design phase. To improve the fault-handling mechanism, we extend the BPEL with two operators, whose semantics are presented. Then the graph planning technology is introduced to validate the fault-handling and compensation-handling mechanisms. The algorithms are detailed respectively and the validation framework is described. The experiment is implemented and the results show that our proposed approach is effective.

For the operator *retry*, we only consider that one service is replaced with another. But, in actual application, one service may be replaced by several services which are combined to satisfy the functional requirement. The loop structure is also not considered in our current solution. Theses will be discussed in our further study. It is limited to guarantee the composition running successfully only with the validation during the design phase in the dynamic environment. So, another part of our future work is to integrate our approach into a self-adaptive framework which can monitor the process execution.

ACKNOWLEDGEMENT

This work is funded by the Natural Science Foundation of China (NSFC, Grant No.60873237), Natural Science Foundation of Beijing (Grant No.4092037), and partially supported by Beijing Key Discipline Program.

REFERENCES

[1] A. A. A. Alves, S. Askary, and et al. April 2007, *OASIS Standard Web Services Business Process Execution Language Version 2.0.* . Available: http://docs.oasis-open.org/wsbpel/2.0/serviceref, [retrieved: March, 2012].

[2] F. Leymann. May 2001, *WSFL: Web Serices Flow Languag*. Available: http://xml.coverpages.org/WSFL-Guide-200110.pdf, [retrieved: March, 2012].

[3] S. Thatte. June 2001, *XLANG: Web Service for Business Process Design.* Available: http://xml.coverpages.org/XLANG-C-200106.html, [retrieved: March, 2012].

[4] M. B. a. C. Ferreira., "An operational semantics for stac,a language for modelling long-running business transactions.," in *Proceedings of Sixth International Conference on Coordination Models and Languages*,, February 2004, pp. 87-104.

[5] Z. Qiu, S. Wang, G. Pu, and X. Zhao, "Semantics of BPEL4WS-Like Fault and Compensation Handling," in *FM 2005: Formal Methods*, ed, 2005, pp. 350-365.

[6] L. Chenguang, Q. Shengchao, and Q. Zongyan, "Verifying BPEL-Like Programs with Hoare Logic," in *TASE '08. 2nd IFIP/IEEE International Symposium on Theoretical Aspects of Software Engineering* 2008, pp. 151-158.

[7] Z. Huibiao, H. Jifeng, L. Jing, and J. P. Bowen, "Algebraic Approach to Linking the Semantics of Web Services," in *SEFM 2007. Fifth IEEE International Conference on Software Engineering and Formal Methods*, 2007, pp. 315-328.

[8] D. Ziyang, A. Bernstein, P. Lewis, and L. Shiyong, "Semantics based verification and synthesis of BPEL4WS abstract processes," in *Proceedings. IEEE International Conference on Web Services*, 2004, pp. 734-737.

[9] X. Fu, T. Bultan, and J. Su, "Analysis of interacting BPEL web services," in *Proceedings of the 13th international conference on World Wide Web*, New York, NY, USA, 2004, pp. 621-630.

[10] M. Xiaoyong, F. Yiyan, W. Yonglin, and F. Xia, "A petri net-based failure handling model for composition transactions," in *Second International Conference onComputational Intelligence and Natural Computing Proceedings (CINC)*, 2010, pp. 378-381.

[11] G. Friedrich, M. Fugini, E. Mussi, B. Pernici, and G. Tagni, "Exception Handling for Repair in Service-Based Processes," *IEEE Transactions on Software Engineering,* vol. 36, pp. 198-215, 2010.

[12] W. Gaaloul, S. Bhiri, and M. Rouached, "Event-Based Design and Runtime Verification of Composite Service Transactional Behavior," *IEEE Transactions on Services Computing,* vol. 3, pp. 32-45, 2010.

[13] Q. L. An Liu, Liusheng Huang,Mingjun, Xiao, "FACTS: A Framework for Fault-Tolerant Composition of Transactional Web Services," in *IEEE Transactions on Services Computing*, 2010, pp. 46-59.

[14] Y. Yan, P. Poizat, and L. Zhao, "Self-Adaptive Service Composition Through Graphplan Repair," in *IEEE International Conference on Web Services (ICWS)*, 2010, pp. 624-627.

[15] C. Tien-Dung, P. Felix, and R. Castanet, "WSOTF: An Automatic Testing Tool for Web Services Composition," in *Fifth International Conference on Internet and Web Applications and Services*, 2010, pp. 7-12.

[16] E. Martinez, M. E. Cambronero, G. Diaz, and V. Valero, "Design and Verification of Web Services Compositions," in *International Conference on Internet and Web Applications and Services* 2009, pp. 395-400.

[17] C. Tien-Dung, R. Castanet, P. Felix, and G. Morales, "Testing of Web Services: Tools and Experiments," in *IEEE Asia-Pacific Services Computing Conference (APSCC)*, 2011, pp. 78-85.

[18] M. L. F. Avrim L. Blum, "Fast Planning Through Planning Graph Analysis," in *Proceedings of the International Joint Conference on Artificial Intelligence(IJCAI)*, 1995, pp. 1636-1642.

A Web Browsing System for Retrieving Scholarly Pages

Ari Pirkola
School of Information Sciences
University of Tampere
Finland
ari.pirkola@uta.fi

Abstract—**The major Web search engines are useful tools to find information from the Web but their commercial nature, coupled with some other factors, makes it difficult to find scholarly pages on a specific topic. Many users prefer browsing to searching because it is easier. Even though browsing subject directories is an important method to retrieve information from the Web, such directories are not suitable for specific scientific retrieval tasks. In this paper, we present a Web browsing system focused on scholarly pages related to a specific topic. The first implementation is dedicated to the topic climate change, but the method to construct the system is a general method that can be applied to any reasonable topic. The climate change browsing system provides access through links to the thematic Web pages of scientific organizations engaged in climate change research, as well as to the pages of organizations that are linked to them. Each link in the system is categorized under a given index term based on the occurrences of phrases related to climate change (keyphrases) on the target pages of the links. In browsing, the user clicks the desired index term and the system returns a list of links to the pages associated with the index term. This paper also presents the crawler used to fetch pages for the browsing system and a keyphrase dictionary used in indexing the pages included in the system.**

Keywords - browsing; climate change; focused crawling; information retrieval; scientific organizations

I. INTRODUCTION

The Web contains tens or even hundreds of billions of documents (pages). Access to its huge content is dominated by commercial search engines, and a concern has been raised about the commercially biased search results. Common experience shows that the pages of companies and other commercial organizations often populate the top ranks of the search results while, for example, many highly informative pages of scientific organizations do not appear in the results. Obviously, this bias is caused by the commercial nature of the search engines and specifically by the search engine optimization (SEO), which means that Web page publishers can promote the ranking of their Web pages in the search results. Commercial organizations are willing and have resources to invest in SEO whereas in the scientific organizations such practice seems not to be common. It has also been discussed where to draw a line between ethical and unethical SEO. The above facts suggest that there is need for such retrieval systems that satisfy the information needs of users who search for scholarly information rather than products or online shops.

The major Web search engines are query-based retrieval systems. Querying is an effective method when the information need can be expressed using a relatively simple query and the search term is clear, e.g., when the searcher looks for information on the disease whose name he or she knows. However, it is common that the searchers do not know or remember the appropriate search terms. Complex information needs are also difficult to formulate as a query. In situations such as complex work tasks the information need itself may be vague and ill defined [1]. Browsing subject directories (e.g., Open Access Directory) is an alternative way to retrieve information from the Web, and it is preferred by many users because it is easier, requiring less mental work than formulating a query. Such directories are created manually and usually in (more or less) unsystematic fashion. Their target pages may be both commercial and informative pages. Even though the subject directories are helpful tools, these features make them unsuitable for specific scientific retrieval tasks. Automatically constructed and systematically organized browsing systems focused on scholarly pages are missing from the Web. In this paper, we present a method to construct such a system.

Using the proposed method we implemented a browsing system (a pilot system) for the topic *climate change*. We present the existing climate change browsing system in this paper, but the proposed method is a general method and can be applied to any reasonable topic. The climate change browsing system contains links to the Web pages of universities, research organizations, and their units (e.g., departments, centers, and institutes) investigating climate change, as well as to the pages of organizations that are linked to them. The system provides information that is difficult or impossible to obtain by the major Web search engines and not included in scientific publications, such as information on current and completed research projects, observational field data and new, not (yet) published research findings. The system allows exploring systematically different aspects of climate change research. In goal-oriented browsing the system provides answers to the questions such as: What organizations conduct research in the field of climate change and what is the specific research area of each organization. Where can I find information related to a specific research area of climate change research? There are also numerous more specific questions where the system may be helpful.

The rest of this paper is organized as follows. Section II presents the main features of the climate change browsing system. The method to construct the browsing system is presented in Section III. Section IV describes the tools needed to construct the system: a focused Web crawler and a keyphrase dictionary of climate change. Section V contains the conclusions.

II. CLIMATE CHANGE BROWSING SYSTEM

In this section, we present the main features of the climate change browsing system. The system provides access to the *thematic pages* of universities, research organizations, and their units, as well as to the thematic pages of organizations that are linked to them (e.g., online journals). Non-thematic pages, such as *about us* and *contact* are pruned out by applying a stop-word list for the URLs of the pages (Section III). We differentiate between two types of thematic pages: *project pages* that describe the research areas of the organizations, or present ongoing or completed research projects, or include some information related to the research projects, and *findings pages* that describe recent research findings (from the present back a few years).

The following simple example illustrates the structure of the browsing system:

melting glaciers 11.7
↓ ↓
Link to page A →
Link to page B → project pages A-C
Link to page C →

Link to page D →
Link to page E → findings pages D-F
Link to page F →

Page titles serve as links, and they are categorized under index terms, such as *melting glaciers*, *climate models*, or *sea level rise* based on the occurrences of the phrases related to climate change (keyphrases) on the pages.

The index terms describe the research areas of the organizations in the field of climate change research, and their source is the climate change keyphrase dictionary (Section IV B). Each index term has an importance score, which reflects the significance of the term in the context of climate change research. As shown in the example above, the index term *melting glaciers* has an importance score of 11.7. The first set of the links (A-C in the example) point to the project pages while the second set (D-F) point to the findings pages. In browsing, the user clicks the desired index term and the system returns the second page through which the user can access the pages discussing the issue represented by the index term.

The first implementation of the browsing system (the pilot system) will be published at [2] in the spring 2012. This site also contains a climate change search system and the keyphrase dictionary, both of which were developed in our earlier project. The search system is described in [3] and the keyphrase identification and extraction method in [4].

III. METHOD

Fig. 1 depicts the crawling and page processing stages involved in constructing the pilot browsing system. In the first stage, relevant pages were crawled for the search system from the scientific Web sites using a focused crawler developed in our earlier project. The crawler is described in Section IV A. Focused crawlers are programs aiming to fetch Web pages that are relevant to a pre-defined domain or topic [5, 6, 7]. The crawled pages were indexed using the Apache Lucene programming library (http://lucene.apache.org/). The constructed search system covers 95 819 (public version 73 194) Web pages.

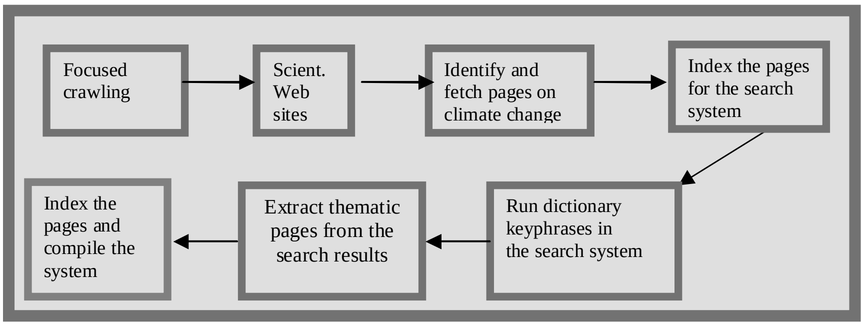

Focused crawling
Scient. Web sites
Identify and fetch pages on climate change
Index the pages for the search system
Run dictionary keyphrases in the search system
Extract thematic pages from the search results
Index the pages and compile the system

Figure 1. The crawling and page processing when constructing the browsing system.

A set of keyphrases contained in the keyphrase dictionary were run as queries in the search system. The purpose of this stage was to retrieve pages dealing the issues represented by the keyphrases. For findings pages, the keyphrases were required to appear in the titles or URLs of pages. In other words, if the title or URL contains the keyphrase, the page is regarded as a findings page. These restrictions were not applied for project pages because their titles and URLs may only be loosely descriptive. If the keyphrase appears in the title or URL we can be highly confident that the page deals with the issue represented by the keyphrase. A drawback of this method is that it excludes part of relevant documents. However, it includes highly relevant documents, which usually are the most important for the user.

The search results were processed to separate the thematic pages from non-thematic pages and to separate the two types of the thematic pages from each other. The identification of the thematic pages is based on the fact that different organizations apply descriptive and similar names for similar items in URLs. Specifically, the URLs of the project pages typically contain the word *research*, *project*, or *science*, e.g., *research-highlights*, *current_projects*, or *research.php*. The URLs of the finding pages in turn often contain the word *news*, sometimes *press* (e.g., *press_release*). We also applied a stop-list of some 50 words indicating non-thematic content (such as *student*, *course*, *contact*, and *grant*) for the URLs to exclude such pages as *www.university.edu/research_grants*.

In the next stage in constructing the browsing system, the URLs of the thematic pages were extracted from the search results, and each page was indexed under the query keyphrase that returned the page (now the keyphrases are called index terms). Finally, the browsing system was compiled by creating appropriate links.

In the keyphrase dictionary, different variant forms of the phrase are grouped together into the same synonym set (Section IV B). Synonymy was taken into account, so that pages containing synonymous keyphrases are under the same index term in the browsing system (i.e., the keyphrase with the highest IS among the synonyms).

IV. THE CRAWLER AND THE KEYPHRASE DICTIONARY

This section presents the crawler used to collect pages for the search system and for source data for the climate change keyphrase dictionary, and describes the dictionary.

A. Crawler

We developed a focused crawler that is used to fetch pages dealing with climate change and that can be easily tuned to fetch pages on other topics. The crawler determines the relevance of the pages during crawling by matching a topic-defining query against the retrieved pages using a search engine. It uses the Lemur search engine (http://www.lemurproject.org/) for this purpose. The pages on climate change were crawled using the following search terms in the topic-defining query: *climate change*, *global warming*, *climatic change*, *research*. We used the core journals in the field to find relevant start URLs. When pages on some other topic are fetched only the search terms and the start URL set need to be changed. So, applying the crawler to a new topic is easy.

To ensure that the crawler fetches mainly scholarly documents its crawling scope is limited, so that it is only allowed to visit the pages on the start URL sites and their subdomains (for example, research.university.edu is a subdomain of www.university.edu), as well as sites that are one link apart from the start domain. If needed, this restriction can be relaxed so that the crawling scope is not limited to these sites.

A focused crawler does not follow all links on a page but it will assess which links to follow to find relevant pages. Our crawler assigns the probability of relevance to an unseen page v using the following formula, which gave the best results in an experiment where we compared four different methods [3]:

$$Pr(T|v) = (\alpha * rel(u) * (1/\log(Nu))) + ((1 - \alpha) * rel(<u,v>)),\ \alpha = 0.3$$

where Pr(T|v) is the probability of relevance of the unseen page v to the topic T, α is a weighting parameter ($0 < \alpha < 1$), rel(u) is the relevance of the seen page u, calculated by Lemur, N_u the number of links on page u, and rel(<u,v>) the relevance of the link between u and the unseen page v. The relevance of the link is calculated by matching the context of the link against the topic query. The context is the anchor text, and the text immediately surrounding the anchor. The context is defined with the help of the Document Object Model (DOM): all text that is within five DOM tree nodes of the link node is considered belonging to the context. The Document Object Model is a convention for representing and interacting with objects in HTML, XHTML and XML documents (http://en.wikipedia.org/wiki/Document_Object_Model).

As can be seen, Pr(T|v) is a sum that consists of two terms: one that depends on the relevance of the page, and one that depends on the relevance of the link. The relative importance of the two terms is determined by the weight α. Based on our crawling experiment we apply for the α parameter the value of α = 0.3. Also, the number of links on page u inversely influences the probability. If rel(u) is high, we can think that the page "recommends" page v. However, if the page also recommends lots of other pages (i.e., N_u is high), we can rely less on the recommendation.

B. Keyphrase Dictionary

To be able to systematically explore a given scientific topic, a necessary requirement is to have a terminology assistance (e.g., ontology, dictionary, thesaurus) that is used to divide the topic into meaningful subtopics or concepts. We constructed the keyphrase dictionary of climate change, which is used to index the organizations for the browsing system. Originally, the dictionary was developed for use as a search assistance to support query formulation in Web searching, but it can be used in document indexing as well. The dictionary contains some 5 500 phrases related to climate change. Most of the phrases represent different aspects or research areas of the climate change research, some are more technical in nature. The phrases were extracted from the Web pages of scientific organizations discussing climate change issues fetched by the crawler described in Section IV A.

Each phrase is assigned a frequency-based *importance score*, which reflects the significance of the phrase in the context of climate change research. Different variant forms of the same phrase, such as *sea level rise*, *sea level rising*, and *rising sea level*, are grouped together into the same entry (synonym set) using approximate string matching.

When devising the dictionary the first challenge was to determine which sequences of words are phrases in the crawled pages. Here we applied the phrase identification method by Jaene and Seelbach [8]. The main point of the technique is that a sequence of two or more consecutive words constitutes a phrase if it is surrounded by small words (such as *the*, *on*, *if*) but do not include a small word (except for *of*).

The other main challenge besides phrase identification was to develop a method to identify the *keyphrases* among all phrases in the relevant pages and prune out out-of-topic phrases. To address this problem, we calculated the importance scores (ISs) for phrases as described below. The most obvious out-of-topic phrases receive a low score and are not accepted in the dictionary. The remaining phrases are regarded as keyphrases and are included in the dictionary.

The IS is calculated on the basis of the frequencies of the phrases in the corpora of various densities of relevant text, and in a non-relevant corpus. We determined the IS using four different corpora. The relevant corpora are built on the basis of the occurrences of the topic title phrase (i.e., climate change) and a few known keyphrases related to climate change in the original corpus crawled from the Web. Assumedly, a phrase which has a high frequency in the relevant corpora and a low frequency in the non-relevant corpus deserves a high score. Therefore, the importance score is calculated as follows:

$$IS(P_i) = \ln(F_{DC(1)}(P_i) * F_{DC(2)}(P_i) * F_{DC(3)}(P_i) / F_{DC(4)}(P_i)); \quad (F_{DC} > 0)$$

$F_{DC(1)}(P_i)$... $F_{DC(4)}(P_i)$ = the frequencies of the phrase P_i in the four corpora.
DC(1) = Highly dense corpus
DC(2) = Very dense corpus
DC(3) = Dense corpus
DC(4) = non-relevant corpus

The presented method allows us to indicate the importance of the phrase in the Web texts discussing the topic in question, and separate between the keyphrases and out-of-topic phrases based on the fact that the relative frequencies of keyphrases decrease as the density decreases.

Synonyms were identified using the digram approximate matching technique. Phrases were first decomposed into digrams, i.e., substrings of two adjacent characters in the phrase (for n-gram matching see [9]). The digrams of the phrase were matched against the digrams of the other phrases in the list of the relevant phrases generated in the phrase identification phase. Similarity between phrases was computed using the Dice formula, and the phrase pairs that had the similarity value higher than the threshold (SIM=0.75) were regarded as synonyms.

Table 1 presents three example entries in the dictionary. The dictionary is organized alphabetically, and each phrase acts as a head phrase in its turn.

V. CONCLUSIONS

In this study, we developed a method to construct a scholarly Web browsing system. The system allows a systematic exploration of a particular scientific topic through browsing the pages of the scientific organizations, and is also a helpful tool in goal-oriented browsing.

Our plan is to extend the existing (but not yet published) pilot browsing system and construct similar systems for new topics. We also want to evaluate the browsing system in a user study. Naturally, many important pages cannot be included in the browsing system if pages are only collected using a crawler. We therefore also plan to implement an upload option for the system, so that the users can suggest and upload URLs to be added to the system.

TABLE I. THREE EXAMPLE ENTRIES IN THE KEYPHRASE DICTIONARY

Head phrase	IS	Synonyms	IS
permafrost thaw	7.3	thawing permafrost	7.9
satellite observation	6.3	satellite observations	11.0
greenhouse gas	25.3	green house gases	7.8
		greenhouse gases	23.9
		greenhouse gasses	13.7
		greenhouses gases	8.1

ACKNOWLEDGMENT

This study was funded by the Academy of Finland (research projects 130760, 218289).

REFERENCES

[1] P. Ingwersen and K. Järvelin. The Tum: Integration of Information Seeking and Retrieval in Context. Heidelberg: Springer, 2005.

[2] http://searchclimatechange.com/

[3] A. Pirkola. "A Web search system focused on climate change." Digital Proceedings, Earth Observation of Global Changes (EOGC). Munich, Germany, April 13-15, 2011.

[4] A. Pirkola. "Constructing topic-specific search keyphrase suggestion tools for Web information retrieval." Proc. of the 12th International Symposium on Information Science (ISI 2011). Hildesheim, Germany, March 9-11, 2011, pp. 172-183.

[5] T. Talvensaari, A. Pirkola, K. Järvelin, M. Juhola, and J. Laurikkala. "Focused Web crawling in the acquisition of comparable corpora." Information Retrieval, 11(5), 2008, pp. 427-445.

[6] S. Chakrabarti, M. van den Berg, and B. Dom. "Focused crawling: a new approach to topic-specific Web resource discovery." Proc. of the Eighth International World Wide Web Conference. Toronto, Canada, May 11-14, 1999.

[7] T. Tang, D. Hawking, N. Craswell, and K. Griffiths. "Focused crawling for both topical relevance and quality of medical information." Proc. of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05). Bremen, Germany, October 31-November 5, 2005, pp. 147-154.

[8] H. Jaene and D. Seelbach. Maschinelle Extraktion von zusammengesetzten Ausdrücken aus englischen Fachtexten. Report ZMD-A-29. Beuth Verlag, Berlin, 1975.

[9] A.M. Robertson and P. Willett. "Applications of n-grams in textual information systems." Journal of Documentation, 54(1), 1998, pp. 48-69.

Optimal Supply Chain Services Management for SMEs through Integrated Model-driven Service System

Catherine Xiaocui Lou
Centre for Strategic Economic Studies
Victoria University
Melbourne, Australia
Xiaocui.lou@live.vu.edu.au

Wei Dai
School of Management and Information Systems
Victoria University
Melbourne, Australia
Wei.Dai@vu.edu.au

***Abstract*— Due to the lack of funding and expertise, small and medium enterprises (SMEs) have largely been excluded from benefiting the spill-over effect of web services-based supply chain systems. Theoretical and empirical researches are at a dearth in that the unmet needs of SMEs have not yet been promoted soundly. This paper assists in filling this gap by contributing to the literature through proposing an integrated model-driven service oriented supply chain framework that makes supply chain for SMEs affordable, easily accessible and free from technical 'hurdles'. The proposed services oriented supply chain system uses a novel framework with five core services coupled with a mathematical optimization model to achieve cost minimization, inventory optimization and reasonable lead time. Simulation results show that our proposed solution is better than the traditional supply chain systems without optimization. Furthermore, it is dynamic and flexible in normal business operation environments. Several simulation run-time examples are used to illustrate the proposed solution.**

Keywords - Optimal model-driven framework; Web Services-based Management Services; supply chain management.

I. INTRODUCTION

Small and Medium Enterprises (SMEs) comprise the largest business segment worldwide. For example, in Europe it is estimated that over 20.7 million SMEs make up 99.8% of all enterprises [1]. Similarly in Australia, they account for 99% of all businesses and are the engine room of the economy. According to the Australian Bureau of Statistics [2], these business entities comprise less than 200 employees and contribute to over 60% of the national employment, innovation, research and development (R&D) and industry value-added [3].

Globally, SMEs generate a huge share of the GDP and are a key resource for new jobs and ongoing employment. They are also a breeding ground for entrepreneurship and new business ideas. As of July 2006, nearly 140 million SMEs around the world employed 65% of the total work force. Apparently, SMEs have been contributed to boosting economic and social development worldwide [4]. Recently, with the advent of online trading, businesses have been able to reach new markets and shorten their supply chains to greatly improve their business connections.

Although there is significant SME participation contributing to the global economy, SMEs are relatively under-represented in the global economy, performing only between one quarter and one third of all manufactured exports, and accounting for less than 10% of Foreign Direct Investment (FDI) [5]. The main barriers that prevent SMEs from being globally active are related to: (1) high cost of hardware and software systems; (2) poor business infrastructure; and (3) inexperienced users of sophisticated software solutions.

As the theoretical and practical literatures of cost-effective and feasible supply chain solutions to SMEs are still sparse [6] [7], this research investigates a service oriented supply chain system. In this, services management operations are integrated with a formal approach in order to address some of the above issues concerning SMEs. The paper is structured as follows. The next section briefly reviewed the literature and our research goal. In Section 3, a supply chain services-based integration framework is proposed and an optimal model is developed for the system, furthermore, solution process associated is provided. Thereafter Section 4 figures out the proposed system architecture and gives out the random simulation example and related results. The last section concludes with the discussions of the limitations and the potential future directions.

II. LITERATURE REVIEW AND RESEARCH GOAL

In contrast to SMEs, the number and scope of successful applications of supply chain operation systems in large companies has grown significantly in recent years. As an illustration, Procter & Gamble drove out non value-adding supply chain costs to save the company over $200 million by using an optimization model with an interactive approach [8]. United Parcel Services (UPS) implemented an optimization modeling system that simultaneously determined aircraft routs, fleet assignments, and package route to ensure overnight delivery at minimal cost. Changes based on a modeling system saved UPS over $87 million between 2000 and 2002 [9].

To tackle the limitations of SME business engagements with its trading partners using the available e-business infrastructure and resources, Supply Chain Management (SCM) has been studied. SCM is about active management of supply chain activities and relationships to maximize customer value and achieve a sustainable competitive advantage [10]. It coordinates all the activities including the materials' physical transformation and information flow from suppliers to the end users. SCM represents a conscious effort by a firm or group of firms to develop and run supply chains in the most effective and efficient ways [11]. NAS defined an integrated supply chain as an association of customers and suppliers who work together to optimize their collective performance in the creation, distribution, and support of an end product (NAS 2000). Zaremba [12] stated that the supply chain aims to be able to link different functions and entities within and outside the company from raw materials to manufacturing, distribution, transportation, warehousing, and product sales. Along the supply chain, a potentially large number of trading partners such as manufacturers, parts suppliers, logistics suppliers, wholesalers and retailers work cooperatively [12]. Furthermore, Charleworth assumed that SMEs are not only seeking ways to integrate the disparate systems within the organization, they also intend to extend the whole domain beyond the boundaries of the organization to include their trading partners and customers [13].

E-business and supply chain applications often involve heterogeneous information resources that may take different standards, protocols and forms and operate in different environments with various complexities. Services-based system platform has advantages in meeting all these potential challenges. Momentum is gathering to apply services system solutions to supply chain problems and proves to be effective [14]. In 2008, Bose, Pal and Ye [15] introduced integration of ERP and SCM systems using the case of a valve manufacturer in China. The improved system successfully reduced lead time and up-grated inventory accuracy. Very recently, To achieve the environmental dynamics, Pan et al. [16] proposed Petri-net-based task model and achieved the connection of low level task with high level system services effectively through the task-to-service mapping algorithm. The research aims to establish a bridge between the arising tasks and potential services to achieve seamless tasks migrations among different application environments. Whilst considering the dynamic cooperation between services system and task framework from Pan et al. [16], our research focuses on the integration of services system and optimal modeling based on the Service Oriented Architecture (SOA). More recently, Choi and Wacker's paper [17] discussed the main theoretical research in the operation management and supply chain management at the aspect of theory building over a period of recent 10 years.

On the other side, in order to help resolve the systematic problems arising in the supply chain process substantially, many researchers are dedicated in the improvement and integration of mathematical models. Huang and Zhen [18] proposed the essential models of the processing in supply chain. In his research, by using the supply chain production strategies with symmetry information, the difference of production strategies under diverse information conditions was analyzed through simulation. However, the reality of producer and consumer determining the production strategies under the asymmetric information condition would cost more at storage and production processing stages. Moreover, for current global network system, not only the producers and stock-keepers relationship as addressed in this research but also the whole supply chain process partners need to be considered. Chang, Wang and Huang [19] studied the cost structure in supply chain. In his research, having the minimum Economic Order Quantity (EOQ) and minimum net profit requirements, the static cost optimization model for distributor was established, with the adjustable parts of customer order quantities as the control variables. Still, the research only discussed some parts of the supply chain operations. In the area of the responsive capacity planning and scheduling, Agrawal, Sephan, and Tsay [20] described a methodology for managing capacity, inventory, and shipments for an assortment of retail products produced by multiple vendors to maximize the retailer's expected gross profit with varied capabilities and demand uncertainty. By systematic examination of the models in SCM research, Narasimban [21] illustrates the five supply chain decision models that demonstrate the importance of integrating the decisions across the SC with their application in global SCM and potential areas. The global economic network also led to the researchers work on global or integrated supply chain models, such as *Huang* [22], Miller [23].

Though there have been well developed researches on the service management and operations optimizations respectively to support SCM, there is few system that successfully combined service management with optimal modeling seamlessly to achieve the real-time integration. To fill this gap, this article illustrates a model-driven based integrated supply chain service framework so that the participants can implement their roles and engagements for efficiency and profitability

III. Solution Approach

A. *The Proposed Model-driven Integrated Supply Chain Services-based Framework*

The main reason behind the SOA adoption is its support for flexible resources allocation, selection and management for SMEs. The model driven approach is to enable a dynamic solution model to match the nature of the tasks. This is to be incorporated into the proposed SOA architecture. The alternatives would be a grid or cloud-based model where the proposed solution model for SMEs would be generated.

Our proposed system incorporates the optimal mathematical modeling into the practical supply chain services management framework through combining both theoretical foundations and business functions within a web services-based system. Further to the research work of Dai and Uden [24], the integrated supply chain service system designed in this paper aims to address the entry barriers of SMEs through the development and provision of core system services that are dynamically integrated with business services to facilitate business operations among trading partners (i.e., consumers and suppliers) in the supply chain. This will require a novel infrastructure in the aspect of integrating formal modeling with supply chain management processes among trading partners for SMEs. To ensure the effectively using of available trading network resources and delivering practical benefits to SME users, our system help SMEs interacting with each other more easily and economically. This is achieved through the integration of global market resources and the real-time communication and interaction support by the proposed service system.

The proposed services oriented supply chain integration has a specialized centre service that coordinates all the businesses including consumer and provider participants within its landscape. The system integrates all the resources within a defined landscape to achieve the goal of optimizing the whole supply chain management through five core services in SOA. These five core services are in the following categories: Knowledge Management Services, Data Management Services, Task Management Services, Information Services and Communication Management Services [24]. With the help of the core services, the higher-level services such as Goal Directed Inference (GDI) service and Event Driven Inference (EDI) service are developed. GDI and EDI services respond to SME users' needs in different ways, e.g. event-driven by triggering purchase order issuing when sales or inventory reaches to a certain level, and goal driven by focusing on user specific request such as fulfilling a specific purchasing request. GDI is particularly supported by two services that are plan generation and plan execution that is supported by the mathematical programming in the next section. GDI provides a model-driven solution in the proposed system. Figure 1 t as attached to this paper shows the technical configuration of the services system.

The participants are supposed to be SME users, who can access the market information in relation to their objectives including low cost and timely delivery through highly optimized and dynamically integrated supply chain channels. The requirements on SME users are to make their consumers requirements for certain product and service in standardized format. The system is to ensure the requirements are transparent to services providers. The process of running the supply chain is executed by the Knowledge Manager (as shown in the Figure 1), which will be improved by the optimization model mentioned in the next section.

B. *Optimal Model-driven Development in the System*

One important contribution towards services oriented supply chain system is to incorporate an optimization model into the service system that includes GDI service. . In order to achieve maximum benefits among the SMEs within the objective supply chain system, a nonlinear optimization model is introduced and described as below.

The annotations for the model are listed as follows.

Sets:

Q : Quantity of the primitive order

Functions or variables in the objective function:

$F(Q)$: Functions for the final integrated supply chains profit;

$f_i(profit)$: Each sub- supply chain profit;

$f_R(price)$: The total income;

$f_P(\cos t)$: The material cost;

$f_c(inventory)$: The inventory expense;

$f_t(transport)$: The transportation fare;

Parameters:

t : The time the whole proposed supply chain process in our system will take;

T_{lim} : Requirement of the time spending;

q_{ij} : Presents the quantity of each independent supply chain.

Optimal Model:

$$F(Q) = Max \sum_{i=1}^{n} \{ f_i(profit) = f_R(price) - f_P(\cos t) - f_c(inventory) - f_t(transport) \}$$

$$subject\ to \quad f_i(profit) \ge \bar{R} \Leftrightarrow Min \frac{1}{2} \alpha \{ f_R(P) - [f_P(C) + f_c(I) + f_t(T)] - \bar{R} \}^2 \text{ ....(1)}$$

$$Min\{ f_P(C),\ f_c(I),\ f_t(T) \} \text{....................................(2)}$$

$$t \le T_{\text{lim}} \text{....................................(3)}$$

$$\sum_{i=1}^{n} \sum_{j=1}^{m} q_{ij} = Q \text{....................................(4)}$$

$$q_{ij} \ge 0 \text{....................................(5)}$$

where Objective function defines the maximum function which including the objective function for the system profits. The part of constraint (1) $Min \frac{1}{2} \alpha \{ f_R(P) - [f_P(C) + f_c(I) + f_t(T)] - \bar{R} \}^2$ is the quadratic penalty function to limit the expectation minimum profit. α is the penalty factor, $\bar{R}$ is set to be company's minimal profit requirement. The second constraint $Min\{ f_P(C), f_c(I), f_t(T) \}$ minimizes the fee of all expenditure therefore to control the cost of the whole supply chain process. The third constraint $t \le T_{\text{lim}}$ is set to meet time requirement from order. Last, q_{ij} presents the quantity of each independent supply chain. The objective is to maximize profit, or reciprocally minimize cost.

To simplify the understanding and usage of the optimal model, we supposed the incoming order with certain price of the productions including unit cost for the inventory, and transportation fare. However, it could certainly be extended to sub-functions for each supply chain processes. For example, the material cost function $f_P(\cos t)$ could be calculated depending on the different proportions of ingredients.

Comparing to the previous supply chain models being used practically, we introduce a penalty function into our optimal system to keep track of the control of the system profit. We also set time constraint to ensure the implement procedure complying with certain required delay time.

C. *Solution Process*

The algorithm of our proposed optimized supply chain is depicted in Figure 2.

Once order requirement comes in, it will trigger the real-time response procedure of our service oriented supply chain system. The basic rules for the system considering the optimal model are, 'first come first serve', 'simultaneous processing of multi-services under time and capacity constraints', and 'check the stock inventory before manufacturing'. Figure 2 describes the main process of the plan generator processing the orders under optimal rules. This algorithm follows up the optimal model described earlier in Section III B.

When the primitive order task entered into the system shown as 'Demand', the procedure is activated. The system will check the inventory based on the 'check the stock inventory before manufacturing' rule. The system then calculates the cost involved and generating production order under the constraints of the model discussed in Section III B.

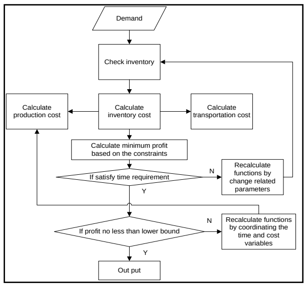

Demand
Check inventory
Calculate production cost
Calculate inventory cost
Calculate transportation cost
Calculate minimum profit based on the constraints
If satisfy time requirement
N
Recalculate functions by change related parameters
Y
If profit no less than lower bound
N
Recalculate functions by coordinating the time and cost variables
Y
Out put

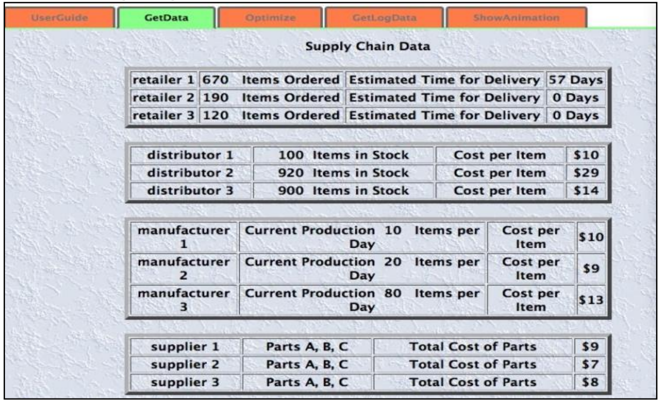
Figure 2 Optimized Model-driven Supply Chain System

IV. SIMULATION AND APPLICATION

The optimization process adopts a multiple objectives, multiple agents approach [25]. The global optimal solution is obtained by mathematical programming. Scenario analysis approach is adopted to illustrate how the proposed system works and compare the performance of the proposed system with the other alternatives.

The processes of one scenario as an example which is simulated by our model-driven based supply chain services system are listed as below. This system is currently under experiment with twelve business entities across four sectors, i.e., retail, distribution, manufacturing and material supply. For run-time simulation, retailers issued purchase orders that trigger dynamic supply chain channels formation. The initial status of simulation in our system is conducted by randomly generated purchase orders from retailers, , e.g., retailer 1 ordered 670 items in figure 3. Secondly, these figures are screen shots that were dynamically taken during the running of our experimental system. Since the initial situation for order is produced randomly by the system, the scenarios could be different when every time you run the experimental system. Last, to help understand our example, it needs to emphasize that our system is working on integrating all the participants' resources of the supply chain and endeavoring to optimize the allocations and chains arrangement. Therefore, the optimization process of our system is trying to work out optimal supply chain solutions for all the participants to save time and cost.

Briefly introducing, the processes for the simulation are divided into: set up the model (including the request, available resources, manufacturing and inventory capability, etc.), e.g., figure 3; calculate the scenario with traditional approach which is not using combination of services management and optimal model-driven, e.g., figure 4; give out our proposed optimal supply chain services management solution, e.g., figure 5.



Supply Chain Data

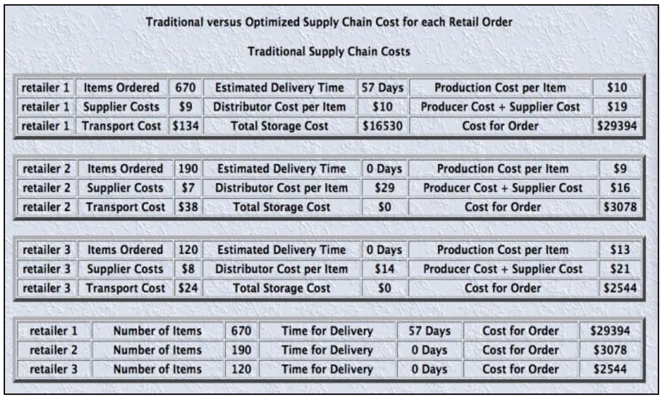
retailer 1	670	Items Ordered	Estimated Time for Delivery	57 Days
retailer 2	190	Items Ordered	Estimated Time for Delivery	0 Days
retailer 3	120	Items Ordered	Estimated Time for Delivery	0 Days

distributor 1	100 Items in Stock	Cost per Item	$10
distributor 2	920 Items in Stock	Cost per Item	$29
distributor 3	900 Items in Stock	Cost per Item	$14

manufacturer 1	Current Production 10 Items per Day	Cost per Item	$10
manufacturer 2	Current Production 20 Items per Day	Cost per Item	$9
manufacturer 3	Current Production 80 Items per Day	Cost per Item	$13

supplier 1	Parts A, B, C	Total Cost of Parts	$9
supplier 2	Parts A, B, C	Total Cost of Parts	$7
supplier 3	Parts A, B, C	Total Cost of Parts	$8

Figure 3 Setting Up the Model



Traditional versus Optimized Supply Chain Cost for each Retail Order

Traditional Supply Chain Costs

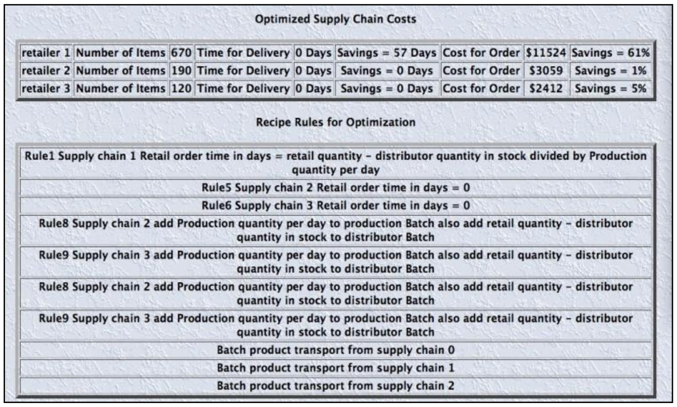
retailer 1	Items Ordered	670	Estimated Delivery Time	57 Days	Production Cost per Item	$10
retailer 1	Supplier Costs	$9	Distributor Cost per Item	$10	Producer Cost + Supplier Cost	$19
retailer 1	Transport Cost	$134	Total Storage Cost	$16530	Cost for Order	$29394

retailer 2	Items Ordered	190	Estimated Delivery Time	0 Days	Production Cost per Item	$9
retailer 2	Supplier Costs	$7	Distributor Cost per Item	$29	Producer Cost + Supplier Cost	$16
retailer 2	Transport Cost	$38	Total Storage Cost	$0	Cost for Order	$3078

retailer 3	Items Ordered	120	Estimated Delivery Time	0 Days	Production Cost per Item	$13
retailer 3	Supplier Costs	$8	Distributor Cost per Item	$14	Producer Cost + Supplier Cost	$21
retailer 3	Transport Cost	$24	Total Storage Cost	$0	Cost for Order	$2544

retailer 1	Number of Items	670	Time for Delivery	57 Days	Cost for Order	$29394
retailer 2	Number of Items	190	Time for Delivery	0 Days	Cost for Order	$3078
retailer 3	Number of Items	120	Time for Delivery	0 Days	Cost for Order	$2544

Figure 4 Traditional Approach to this Scenario



Optimized Supply Chain Costs

retailer 1	Number of Items	670	Time for Delivery	0 Days	Savings = 57 Days	Cost for Order	$11524	Savings = 61%
retailer 2	Number of Items	190	Time for Delivery	0 Days	Savings = 0 Days	Cost for Order	$3059	Savings = 1%
retailer 3	Number of Items	120	Time for Delivery	0 Days	Savings = 0 Days	Cost for Order	$2412	Savings = 5%

Recipe Rules for Optimization

Rule1 Supply chain 1 Retail order time in days = retail quantity – distributor quantity in stock divided by Production quantity per day
Rule5 Supply chain 2 Retail order time in days = 0
Rule6 Supply chain 3 Retail order time in days = 0
Rule8 Supply chain 2 add Production quantity per day to production Batch also add retail quantity – distributor quantity in stock to distributor Batch
Rule9 Supply chain 3 add Production quantity per day to production Batch also add retail quantity – distributor quantity in stock to distributor Batch
Rule8 Supply chain 2 add Production quantity per day to production Batch also add retail quantity – distributor quantity in stock to distributor Batch
Rule9 Supply chain 3 add Production quantity per day to production Batch also add retail quantity – distributor quantity in stock to distributor Batch
Batch product transport from supply chain 0
Batch product transport from supply chain 1
Batch product transport from supply chain 2

Figure 5 Proposed optimal supply chain services management solution

The result shows that for the scenario absent of supply chain integration, the three different supply chains' total cost adds up to $35016 and the lead time is 57 days. While through using our proposed model in this paper, the total cost is reduced to $16995, which saves more than 50% overall.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

Though the savings in the costs are disproportionately distributed, it is apparently that all the SMEs are better off. The lead-time reduced to 0, which is consistent with the “just-in-time” approach.

SME users can conveniently access our new integrated model-driven service through a Web browser or hand-held devices such as mobile phones from any corner of the world. An operational system configuration can be found in Figure 6.

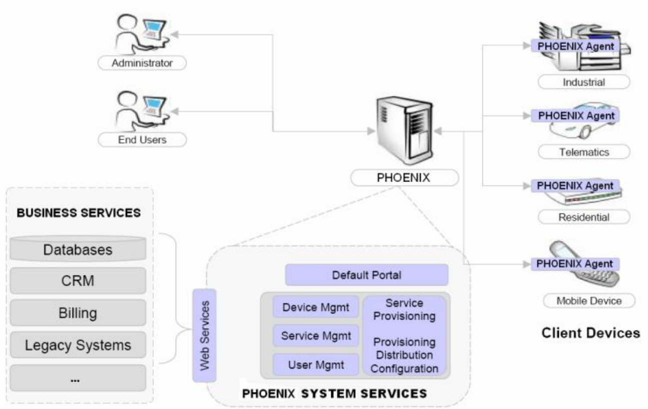


Figure 6 The PHOENIX Services System Architecture

V. Conclusions And Future Work

The research proposed a feasible and cost-effective way to enable SMEs to access supply chain management tools, which used to be the privilege of large corporations. The marriage between supply chain integration and mathematical optimization techniques is a critical contribution to innovation by this research. The simulation results show that our proposed framework significantly improves SMEs’ situation by saving costs and reducing the lead time.

This paper is subjected to the following limitations: (1) the simulation is not robust enough to produce any general conclusions; (2) minutes of all the details within the supply chain has yet to be specified; (3) the input and output communication among all the parties have not been considered.

Future research will focus on applying the proposed framework to the real world situations. In addition, a mobile set- based model can also be designed to free users from all the details of supply chain and arduous supply chain management activities.

Acknowledgment

We would like to gratefully acknowledge the support from the Centre for Strategic Economic Studies in Victoria University and School of Engineering and Information of Beihang University (BUAA). Particularly give our sincerely appreciate to Professor Qishan Zhang in Beihang University (BUAA), who gave us indelible guidance. Thanks Mr. Paul Moynahan, who developed a prototype Web service demonstration system at Victoria University, and Mr. Yongqiang Li in discussion and helping polish up this article.

References

[1] W. Paul, S. Viera, D. James, and A. Barke, “Are EU SMEs Recovering?”, Annual Report on EU SMEs. European Commission, 2010/2011.

[2] ABS, “Small business in Australia,” Cat. no. 1321.0. 2002.

[3] A. Armstrong, A. Clarke, Y. Li, and K. Heenetigala, “Developing a Responsive Regulatory System for Small Business: Governance in Small Business,” ISBN 978-1-8628-72-692-5. Melbourne, Victoria University, 2011.

[4] European Commission (2003 a), “The New SME Definition: User Guide and Model Declaration,”

http://ec.europa.eu/enterprise/policies/sme/files/sme_definition/sme_user_guide_en.pdf, [retrieved: April, 2012]

[5] C. Hall, “Profile of SMEs and SME Issues in APEC,” for the APEC Small and Medium Enterprises Working Group in Cooperation with PECC (Pacific Economic Cooperation Council), 1990-2000.

[6] J. Campbell and J. Sankaran, “An Inductive Framework for Enhancing Supply Chain Integration,” International Journal of Production Research, 43(16): 3321-3351, 2005.

[7] X. Lou and W. Dai, “Supply Chain Integration and Optimization Model for Small and Medium Enterprises (SMEs),” Recent Achievement on Merging Supply Chain and E-Commerce in China: 258-265. 2009.

[8] J. Camm, T. Chormna, F. Dill, J. Evans, D. Sweeney, and G. Wegryn, “Blending OR/MS Judgment and GIS: Restructuring P&G’s Supply Chain,” Interfaces, Vol. 27, 1997, pp. 128-142.

[9] A. Armacost, P.C. Barnhart, K.A. Ware, and A.M. Wilson, “UPS Optimizes Its Air Network,” Interfaces, Vol. 34, 2004, pp. 15-25.

[10] Robert B. Handfield and Ernest L. Nichols, “Introduction to Supply Chain Management,” Prentice Hall, Upper Saddle River, NJ, 1992.

[11] Cecil C. Bozarth and Robert B. Handfield, “Introduction to Operations and Supply Chain Management,” second edition, Pearson Education, 2008.

[12] M. Zaremba, S. Zaleski, B. Wall, and J. Browne, “Internet Enabled Supply Chain Integration for SMEs,” 2003.

http://csrc.lse.ac.uk/asp/aspecis/20030182.pdf, [retrieved: April,

2012]

[13] I. Charleworth, J. Hamilton, M. Holden, E. Holt, T. Jagger, T. Jennings, and T. Jones, “EAI and Web Services: Cutting the Cost of Enterprise Integration,” in Technology, 2002.

[14] S. Kumar, V. Dakshinamoorthy, and M.S. Krishnan, “Does SOA Improve the Supply Chain? An Empirical Analysis of the Impact of SOA Adoption on Electronic Supply Chain Performance,” Proceedings of the 40th Hawaii International Conference on System Sciences, IEEE Computer Society Press, 2007.

[15] Bose Indranil, Pal Raktim, and Ye Alex, “ERP and SCM Sytems Integration: The Case of a Valve Manufacturer in China,” Information & Management, Vol. 45, 2008, pp. 233 - 241.

[16] G. Pan, Y. Xu, Z. Wu, S. Li, L.T. Yang, M. Lin, and Z. Liu, “Task Shadow: Toward Seamless Task Migration across Smart Environments,” IEEE Intelligent Systems, May/June Issue, IEEE Computer Society Press, 2011, pp. 50 -57.

[17] Thomas Y. Choi and John G. Wacker, "Theory Building in the OM/SCM Field: Pointing to the Future by Looking at the Past," Journal of Supply Chain Management, Vol. 47, No. 2, 2011, pp. 8-11.

[18] X. Huang and L. Zhen, "Production Strategy in Supply Chain under Asymmetric Information," Chinese Journal of Management Science, Vol. 10, No. 2, Apr. 2002:35-40

[19] L. Chang, J. Wang, and X. Huang, "The Cost Model and Its Optimization in Supply Chain," System Engineering, Vol. 20, No. 6, 2002.

[20] Narendra Agrawal, Stephan A. Smith, and Andy A. Tsay, "Multi-vendor Sourcing in a Retail Supply Chain," Production and Operations Management, Vol. 11, No. 2, 2002, pp. 157-82.

[21] R. Narasimhan and S. Mahapatra, "Decision Models in Global Supply Chain Management," Industrial Marketing Management, Vol. 33, No. 1, 2004, pp. 21-7.

[22] George Q. Huang, X.Y. Zhang, and L. Liang, "Towards Integrated Optimal Configuration of Platform Products, Manufacturing Processes, and Supply Chains," Journals of Operations Management, Vol. 23, 2005, pp. 267-290.

[23] Miller Tan and Matta Renato de, "A Global Supply Chain Profit Maximization and Transfer Pricing Model," Journal of Business Logistics, Vol. 29, No.1, 2008.

[24] W. Dai and L. Uden, "Empowering SME Users through Technology Innovation: A Services Computing Approach," Journal of Information and Knowledge Management. World Scientific Publishing, Vol. 7, No. 4, 2008, pp. 267-278.

[25] R.B. Chase and F.R. Jacobs, "Operations Management for Competitive Advantage," McGraw-Hill/Irwin Series, Operations and Decision Sciences 11th, 2006.

ATTACHMENT

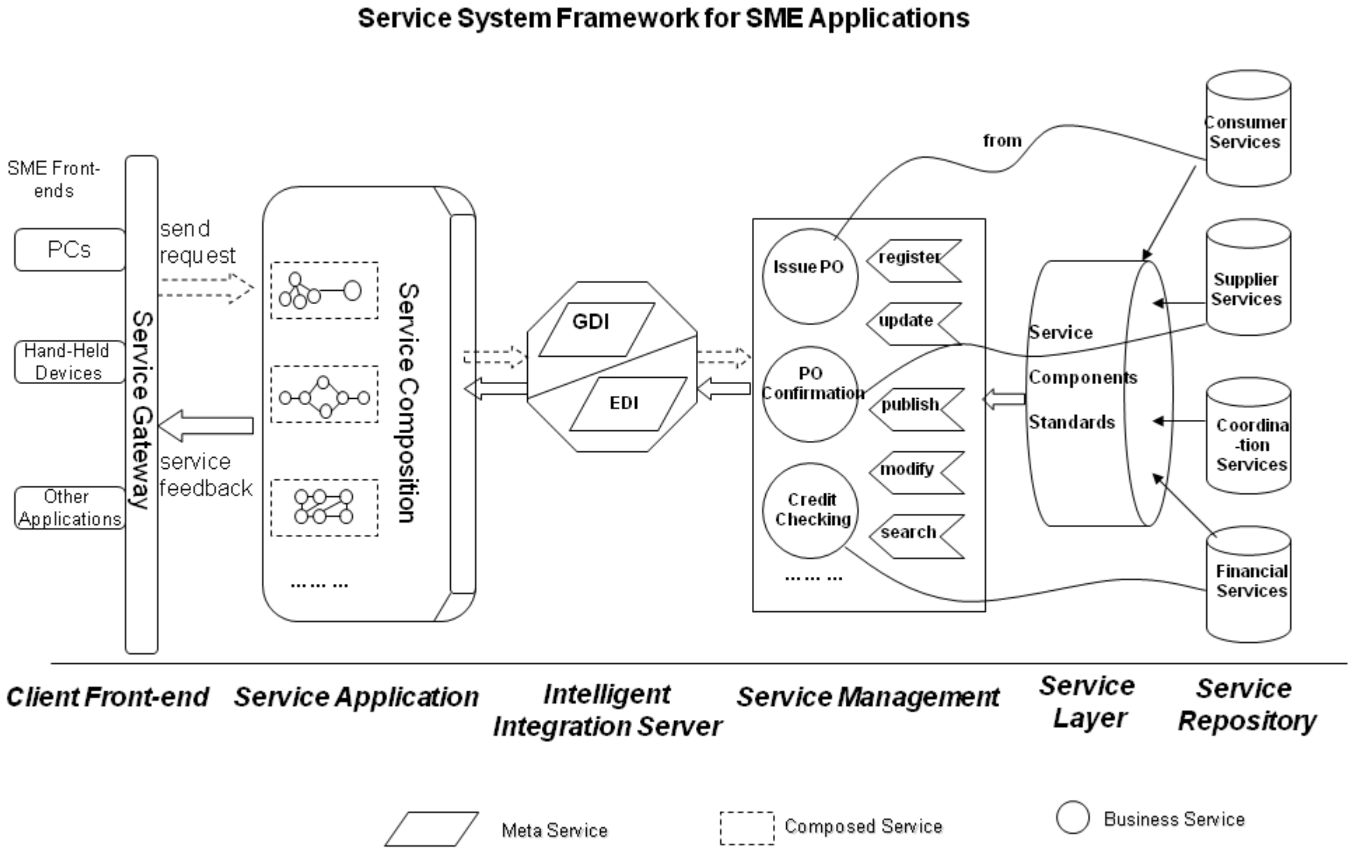


Figure 1 Services Oriented System Architecture

Reliability and Message Security for Distributed Web Service Handlers

Beytullah Yildiz
Department of Computer Engineering
TOBB Economics and Technology University
Ankara, Turkey
E-mail: byildiz@etu.edu.tr

***Abstract*—Web Service handlers are supportive functionalities and capabilities to the service endpoint such as security, reliability and logging. In the common usage, they perform their executions in a single memory space with the service endpoint. However, by a suitable structure, they can be distributed to increase availability, scalability and performance. On the other hand, the distribution necessitates additional mechanisms to provide essential service quality. In this paper, reliability and message security for the distributed Web Service handler will be investigated. The benchmark results are provided to illustrate that the utilized reliability and security mechanisms of the messaging for the handler distribution are reasonable. With the fair cost, Web Service handlers are reliably and securely distributed.**

Keywords-Web Service; distributed computing; replication; reliability; security

I. INTRODUCTION

Web Service is a technology providing seamless and loosely coupled interactions, which help to build platform independent distributed systems. Web Service is considered to be an ideal technology to provide new IT architectures. It is claimed that the age of proprietary information systems has come to an end and the age of shared services is already on its way [1]. In this new era, companies obtain or outsource their IT capabilities in order to reduce the cost, deploy solutions faster, and create new opportunities.

Software standards and communication protocols providing the common languages are at the foundation of Web Service. Information is easily exchanged between different applications via these standards and protocols. In short, Web Service provides opportunities so that diverse and distributed applications can communicate with each other in a standard way.

Web Service integrates an endpoint and handlers in a common framework. It employs supportive functionalities and capabilities, called as Web Service handlers, to provide a full-fledged service. These capabilities might be related to security, reliability, orchestration, logging as well as any necessary capabilities for a distributed system. A Web Service may employ several handlers in a single interaction. In other words, a chain of handlers can contribute to a service execution. With these additive functionalities, Web Services aim to offer better environment. On the other hand, overloading a service with required supportive functionalities, inevitable for many cases, may cause degradation in the service quality. A service endpoint with many handlers may suffocate in a single memory space. Hence, it is wise to use additional computing power. This brings the idea of distribution. There are different reasonable approaches for the Web Service handlers for the distribution. Some suggest that they can be distributed as services; others create a specific distributed environment for them. By creating a specific environment, a distributed handler operating system provides a better environment, especially, when the concern is performance. However, the distribution requires additional mechanisms to provide the suitable environment.

Security and reliability are among the most important criteria that need to be considered when a distributed system is being evaluated. This paper investigates reliability and message security for the distributed Web Service handlers and their effect over the system performance. The rest of this paper is organized as follows. Section II provides information about the related works of reliability and security. Distributed Web Service handler execution is briefly explained in Section III. Section IV investigates reliability. Section V gives details about the message security. Finally, the paper will be concluded in Section VI.

II. RELATED WORKS

Reliability and security are very important for the distributed applications. Many researches on security and reliability have been conducted for the distributed applications in [2] [3] [4] [5].

For Web Services, several standards are provided for the security and reliability purpose: WS-Security [6], WS-ReliableMessaging [7]. WS-Security addresses security by leveraging existing standards and provides a framework to imbue these mechanisms into a SOAP message. This happens in a transport-neutral fashion. WS-Security defines a SOAP header element to carry security related data. This header element contains the information defined by XML signature that conveys how the message was signed, the key that was used, and the resulting signature value. Likewise, the encryption information can be inserted to the SOAP header. In short, WS-Security presents an end-to-end solution for Web Service security by keeping all security information in the related SOAP header element.

ISBN: 978-1-61208-200-4

The WS-ReliableMessaging specification offers an outline to ensure reliable message delivery between the sender and receiver. The specification provides an acknowledgement based scheme to guarantee that data are transferred between the communicating entities. Although it is for the point-to-point communication, the specification also supports service composition and transactional interaction.

III. DISTRIBUTION

Web Service handlers are executable in the distributed environment to meet necessary requirements and to provide enough computing power for Web Services. Distribution improves scalability, availability and performance of the overall system. On the other hand, it brings challenges. A manager for the distributed Web Service handlers must be employed to organize the execution which contains an orchestration mechanism, explained in [8]. The manager also requires a decent execution engine to meet the performance requirements. The details of the manager are provided in [9]. The distribution overhead must be acceptable, which is investigated in [10].

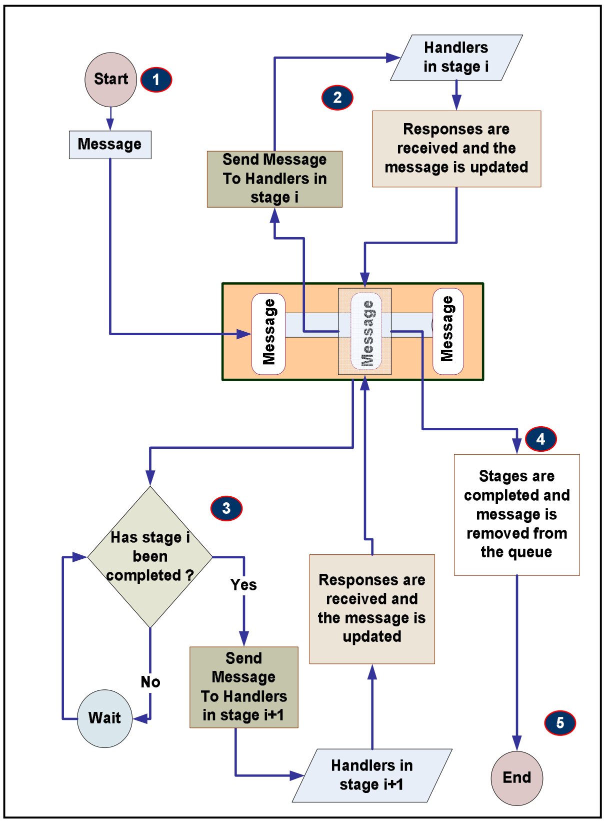

Start
1
Message
2
Handlers in stage i
Send Message To Handlers in stage i
Responses are received and the message is updated
Message
Message
Message
4
Stages are completed and message is removed from the queue
3
Has stage i been completed ?
Yes
No
Wait
Send Message To Handlers in stage i+1
Responses are received and the message is updated
Handlers in stage i+1
5
End

Figure 1. Executing the messages in the distributed Web Service handlers.

The execution of the messages is shown in the Figure 1. Messages, stored in a processing queue, are executed concurrently. Manager ensures that each message is executed without being interrupted by the remaining messages in the queue. Every message execution contains stages, which host the distributed handlers.

A message in the processing queue is instantly sent to all handlers of a stage. The handlers in a stage are executed in a parallel manner. The manager waits the completion of the handler executions before starting the delivery of the message to the next stage. This procedure continues until all stages of a message are completed. In this process, since handlers are deployed to the remote machines, the security and reliability of the messaging become important. The reliability of a handler itself is also essential for the successful execution.

IV. RELIABILITY

Software reliability is described as the probability that the software functions without failures under given conditions during a specified period of time [11]. Reliability is also measured in terms of percentage of failure circumstances in a given number of attempts to compensate for variations in usage over time [12]. For Web Services, although reliability is viewed by some researchers as a non-functional characteristic [13], Zhang and Zhang describes one of the more comprehensive definitions of Web Services reliability, which is defined as a combination of correctness, fault tolerance, availability, performance, and interoperability, where both functional and non-functional components are considered [14].

In this paper, the reliability will be investigated in two sections: the reliability originating from the handler replication and the reliability coming from the utilization of a reliable messaging system.

A. *Replicating handlers*

Replication is critical to reliability, mobility, availability, and performance of a computing system. We benefit from the replication in our daily life too. Even our body benefits from the replications; we have two legs, hands, eyes and ears. We keep a spare tire in our car to replace a flat one in an emergency. The important files are backed up to reduce the probability of lost. Software systems also utilize the same strategy by replicating the data and the computing nodes.

There are basically three replications: data, process and message. These concepts are extensively explored in [15]. Data replication is the most heavily investigated one. However, the other replications are also very important in the distributed systems, especially for Service Oriented Architecture.

The process replication is particularly main interest in this paper because the intention is to investigate the replication of the handlers. There exist two main approaches in this area. The first one is *modular redundancy* [16]. The second approach is called *primary/standby* [17]. *Modular redundancy* has the replicated components that perform the same functionalities. All the replicas are active. On the other hand, *primary/standby* approach utilizes a primary replica to perform the execution. The remaining replicas wait in their standby state. They become active when the primary replica fails.

The processes can be classified in two categories; no consistency and consistency. The fist category is the simplest one; the processes are stateless. They do not keep any information for the processed data. Therefore, the consistency is not an issue between the processes. Replicated instances can be allowed running concurrently. On the other hand, replicas may enter in an inconsistent state if the process is not atomic and statefull. Inconsistency have been extensively investigated in [18].

Replication is a very important capability where a handler is inadequate. Sometimes, a handler may not be sufficient to answer the incoming requests. The tasks may line up so that the overall performance degrades. This is similar to a shopping center where the customers are waiting in the line to be served. The solution is to add one more person to serve when it is necessary. Similarly, adding a handler to help the execution contributes the overall performance.

In addition to the performance, a replica can be leveraged for fault tolerance. It is possible that a handler crashes. The replication contributes to the continuity of the execution and improves availability and reliability of the service. Without using handler replication in the case of an error, the whole computation cannot continue. The computation becomes more resilient with the handler replication. The execution continues while at least one replica of every handler has not failed.

For N handlers with the replication factor of R, the execution can be successful for R-1 failures per handler. The maximum allowable number of error is:

$$\sum_{i=1}^{N} R_i - 1 \qquad (1)$$

where N is the number of handlers, R_i is the replication number of ith handler. The system cannot continue its execution even in a single handler fault where$\forall i \in N: R_i = 1$.

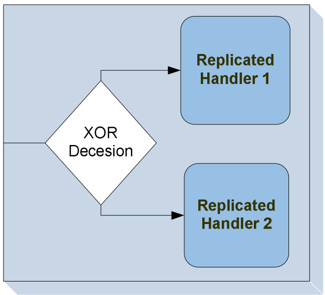


Figure 2. Replicated handler execution; only one of the handlers can be executed.

In the distributed Web Service handler execution environment, a variation of *primary/standby* approach is utilized. The replicas are prioritized. The handler having highest priority is assigned to execute a message. The other replicas wait until their priorities become highest. The system is able to change the priority during the execution. When a fault occurs, the handler priority is minimized. The replicas are never allowed to be executed concurrently unless they are the instance of the stateless handlers. Even though they are allowed to run in parallel manner, they cannot join the processing of the same message. The messages have to be different so that the parallel execution does not cause inconsistency.

When only one of the several replicated handlers is executed, shown in Figure 2, the following formula works for the reliability:

$$R_{RH} = \sum_{i=1}^{n} P_i\, R_i \qquad (2)$$

where R_{RH} is the reliability of the replicated handlers' execution, P_i is the execution probability of the handler i and $\sum_{i=1}^{n} P_i = 1$

The reliability of parallel handlers with AND junction and the reliability of serial handlers can be formulated as:

$$R_s = \prod_{i=1}^{n} R_i \qquad (3)$$

where $R_s$is the reliability of the handlers’ execution and R_i is the reliability of the handler i.

By using Formulas 2 and 3, the reliability of handlers' execution in Figure 3 can be formulated as:

$$R_s = \prod_{i=1}^{2} R_i * \sum_{Ri=1}^{2} P_{Ri} R_{Ri} \;\; * R_3 \qquad (4)$$

$$R_s = \prod_{i=1}^{3} R_i * \sum_{Ri=1}^{2} P_{Ri}\, R_{Ri} \qquad (5)$$

where R_s is the reliability of handlers’ execution. Ri is the ith replica and $P_{Ri} = 1$ for only one replicated handler, which is executed, and the value is 0 for the remainders.

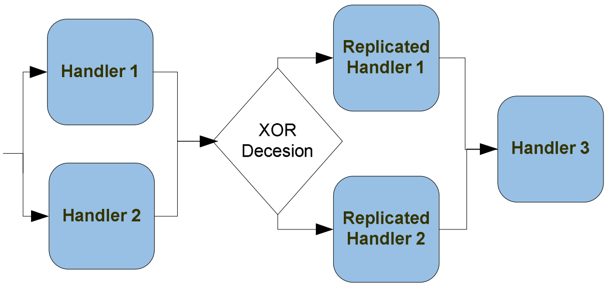


Figure 3. A sample configuration for the handlers' excution

B. Reliable messaging

The distributed handler mechanism benefits from two different sources for the reliability of the message delivery: a messaging broker, and its own mechanism.

The messaging system, NaradaBrokering, provides a message level reliability. It also offers supportive functionalities for the messaging and grants very reasonable performance [19]. The messages can be queued up to several thousands and are gradually delivered to their destinations to provide a flow control for the messaging. Additionally, it has Reliable Delivery Service (RDS) component that delivers payload even if a node fails [20].

RDS stores all the published events that match up with any one of its managed templates, which contain the set of headers and content descriptors. This archival operation is the initiator for any error correction, which is caused by the events being lost in transit to their targeted destinations and also by the entities recovering either from disconnect or a failure. For every managed template, RDS also maintains a list of entities for which it facilitates reliable delivery. RDS may also manage information regarding access controls, authorizations and credentials of the entities that generate or consume events, which are targeted to this managed template.

When an entity is ready to start publishing events on a given template, it issues a discovery request to find out the availability of RDS that provides archival environment for the generated template events. The publisher will not circulate template events until such time that it receives a confirmation that RDS is available.

The publisher ensures that the events are stored by RDS for every template event that it produces. After successful delivery of the event to RDS, it is archived and a message is sent to the publisher to verify that the message is received by RDS successfully. Otherwise, a message of failure with the related event id is sent back to the publisher. After having the verification, the suitable matching engine is utilized to compute the destinations associated with the template event.

A subscriber registers with RDS. A sequence number linked with the archival of this interaction is recorded. The number can be also described as epoch, which signifies the point from which the registered entity is authorized to receive events conforming to the template. Once a template event has been archived, RDS issues a notification. The notifications allow a subscribing entity to keep track of the template events while facilitating error detection and correction. Upon receipt of the notification, the subscribing entity confirms the reception of the corresponding template event.

When an entity reconnects to the broker network after failures, the entity retrieves the template events that were issued and those that were in transit before the entity leaving. After the receipt of the recovery request, RDS scans the dissemination table starting at the sync related with the entity and then generates an acknowledgment-response invoice event outlining the archival sequences, which the entity did not previously receive. Accordingly, the missing events are provided to the receiver.

In addition to this, a reliable mechanism for Web Service handler execution environment is built on the top of the reliable messaging that NaradaBrokering provides. The distributed Web Service handler mechanism is able to repeat the execution of a specific handler in the situation of a failure. The decision of a failure is made when the response is not received from a distributed handler. There can be several reasons behind being unsuccessful to get a response. The communication link may be broken as well as the handler may not successfully process the message because of either an error or crash. The distributed Web Service handler mechanism checks the possibilities by sending the message several times to its destination. In each attempt, it waits for a specific amount of time. This duration is either assigned or calculated by the system. After having several unsuccessful attempts, the message processing may switch to a replica if it exists. As it is discussed previously, handlers can populate their replicas to improve availability and reliability.

For the reliable messaging benchmark, two HP DL 380 G7, 2 x Xeon Six Core, 2.93 GHz, and 48 GB memory physical machines are utilized. The machines are virtualized to create four 4-core and 16 GB memory machines and one 8-core 32 GB memory machine. These machines are connected to each other via LAN and share a common storage system. Virtual machines use Windows Server 2008 R2 64-bit operating systems. The cost of reliable mechanism of the messaging for the distributed handlers is shown in Figure 4. The cost contains the time of reliability procedures to send the tasks to the distributed Web Service handlers or receive the responses back. The time for the handlers' executions and the time for the messaging are excluded to illustrate only the reliability cost for varying message sizes. The figure shows that the message size does not affect the cost of the reliability of the messaging very much. The cost is very reasonable when the reliability is a necessity for the distribution.

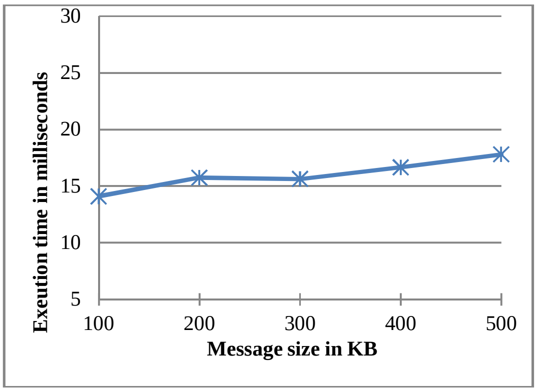


Figure 4. The cost of reliability mechanism of the messaging for the distributed handlers

V. Message Security

Security is one of the important issues for the computing systems. The very critical data can be seen or altered by an unauthorized person. This is increasingly important if the data is transferred through the network, which is more vulnerable environment.

The local computing is not exposing its data to the outside world very much. In contrast, this is not the case for the distributed computing. The computation is shared between the nodes which may physically disperse in the distributed environment. The transmission of the data among the nodes may expose the critical information to the dangerous vulnerabilities. Hence, the transportation channels must be secured in addition to the security of the computing entities.

NaradaBrokering, which is utilized for messaging, has a security framework that is able to support secure interactions between the distributed handlers [21]. The security infrastructure consists of Key Management Center (KMC), which provides a host of functions specific to the management of keys in the system. At the same time, KMC incorporates with an authorization module to manage the usage of the messaging. KMC also stores the entities public keys.

NaradaBrokering has an authentication mechanism for the publishers and subscribers, which are the computing nodes for the distributed handler execution. For the authentication, publisher or subscriber sends its signed request by using private key. Every topic has access control list which authorizes the subscriber. Similarly, an access control list exists for the publishers. After verification of signature, entity is permitted to be accessed by the publisher or subscriber according to the relevant access control lists.

The message traveling between the computing nodes is described in Figure 5. It contains a unique id, properties and a payload. Unique message id is a distinctive name for a message. The handler execution mechanism may host many messages being executed in a moment. Hence, an identifier is a necessity to achieve the correct executions; a Universally Unique Identifier (UUID) generated id is assigned to every message. The generator assures that there won't be the same id in the system. Thus, the design gives enough guarantees that the message executions are not blended.

```
<context>
  <id>4099d6dc-0b0e-4aaa-95ff-2e758722a959</id>
    <properties>
     <encKey>abcdef</encKey >
      ….
  </properties>
  <payload>
    ….
  </payload>
</context>
```

Figure 5. The message format for distributed Web Service handlers

The second important part of the message format is the properties section. This part conveys the required additional information for the computing nodes. The information can be specific to a handler as well as generic for all handlers. There is a property that contains a key for the encryption. It is a session key which is created for a single message. However, the session key can be utilized to send a group of messages to a distributed handler for a period of time. The payload containing the original message is encrypted by this key before sending to its destination to keep the message integrity intact.

In many distributed design, secure data transmission is not discussed, the models rely on the existing security technology such as Secure Socket Layer (SSL). Kemathy at al. investigates component base solution for XML messaging [22]. Ammari at al. provides architecture securing XML messages by encrypting flagged XML parts each with different type of encryption depending on data sensitivity and importance level defined [23], Figure 6 demonstrates the secure messaging for the distributed handlers. The XML based message of the distributed handlers is partially encrypted, only the payload. Since encryption via asymmetric key performance is worse than the symmetric key encryption [24], Advanced Encryption Standard (AES) symmetric key encryption algorithm is used to encrypt the payload. A 256 bit session key is created for each message and passed within the message to the other computing node for decryption. The sender encrypts the session key with the 2048-bit public key of the receiver to present the confidentiality. The related public key is provided by the KMC. RSA algorithm is used for the key encryption. Hence the only node, which has the correct private key, can decrypt the session key to get the payload.

When the subscriber receives the message, first of all, it decrypts the session key carried within "encKey" tag with its private key. Then, the session key is used to decrypt the payload to get the original message.

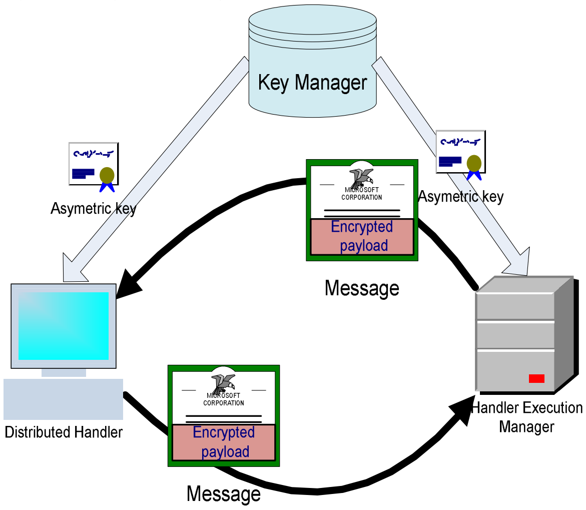


Figure 6. Security mechanism for a distributed handler

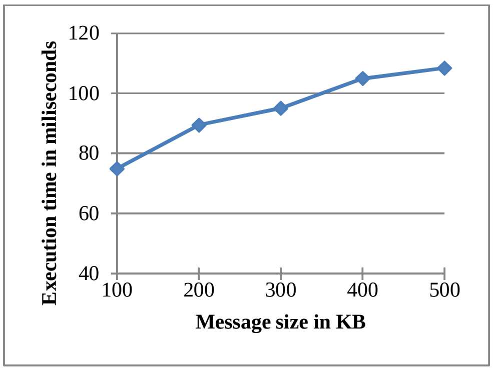


Figure 7. The cost of the security mechanism of the messaging for the distributed handlers

The benchmark showing the cost of the aforementioned security mechanism for the messaging is performed in the same environment with the reliability benchmark, discussed Section IV.B. Figure 7 shows the cost for varying payload sizes. The usage of the symmetric key encryption provides reasonable execution time. Even though the reliability offers better results, the cost of security does not grow exponentially for the increasing message size.

VI. CONCLUSION

While the distribution of Web Service handlers provides many advantages in terms of scalability, availability and performance, the environment necessitates reliability and secure messaging. The instruments, explained in this paper, for the secure and reliable handler distribution and the support tools of the utilized messaging broker grant the necessary reliability and messaging security for this environment. The benchmark results show that the costs originating from the utilized instruments are acceptable. The replication of the handlers contributes the execution during failures. In short, the design of the distributed execution with the security and reliability offers a satisfactory environment for Web Service handlers.

REFERENCES

[1] J. Hagel and J.S. Brown, “Your next IT strategy,” Harvard Business Review, 79 (10), pp. 105-113, 2001.

[2] P. Bzoch and J. Safarik, "Security and reliability of distributed file systems," in IEEE 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS 2011). vol.2, pp.764-769, Sept. 2011, doi: 10.1109/IDAACS.2011.6072873

[3] M. Lei, S. V. Vrbsky, and Z. Qi, “Online grid replication optimizers to improve system reliability,” in IEEE International Parallel and Distributed Processing Symposium (IPDPS 2007), pp.1-8, March 2007.

[4] K. Birman, R van Renesse, and W Vogels, "Spinglass: secure and scalable communications tools for mission-critical Computing," in International Survivability Conference and Exposition (DARPA DISCEX-2001), CA, June 2001.

[5] C. M. Jayalath and R. U. Fernando. “A modular architecture for secure and reliable distributed communication,” in Proceedings of the Second International Conference on Availability, Reliability and Security (ARES07), pp. 621-628, 2007, Washington, DC. DOI=10.1109/ARES.2007.7 http://dx.doi.org/10.1109/ARES.2007.7

[6] Web Service Security (WS-Security), http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wss/, <retrieved: 03, 2012>.

[7] Web Service Reliable Messaging (WS-ReliableMessaging), http://public.dhe.ibm.com/software/dw/specs/ws-rm/ws-reliablemessaging200502.pdf, <retrieved: 03, 2012>.

[8] B. Yildiz, G. Fox, and S. Pallickara, “An orchestration for distributed Web wervice handlers,” in International Conference on Internet and Web Applications and Services (ICIW08), pp. 638-643, June 2008, Athens, Greece

[9] B. Yildiz, “Distributed handler architecture,” Ph.D. Dissertation. Indiana University, Bloomington, IN, USA. Advisor: Geoffrey C. Fox. 2007.

[10] B. Yildiz and G. Fox, "Measuring overhead for distributed Web Service handler," in Proceedings of 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT 2010), pp. 566-570, July 2010.

[11] H. Zo, D. Nazareth, and H. Jain, "Measuring reliability of applications composed of Web Services," in Proceedings of 40th Annual Hawaii International Conference on System Sciences (HICSS '07), pp. 278- 288, 2007.

[12] J. D. Musa, "Software reliability engineering," McGraw-Hill, New York, NY, 1999.

[13] A. Arsanjani, B. Hailpern, J. Martin, and P. Tarr, “Web Services: promises and compromises, ”ACM Queue, 1 (1), pp. 48-58, March 2003.

[14] J. Zhang and L.-J. Zhang, “Criteria analysis and validation of the reliability of Web Services-oriented systems,” in Proceedings of the IEEE International Conference on Web Services (ICWS'05), Orlando, Florida, July 2005.

[15] A. Helal, A. Heddaya, and B.K. Bhargava, "Replication techniques in distributed systems," Kluwer Academic Pub. 2002, Volume 4, pp. 61-71, DOI: 10.1007/0-306-47796-3_3.

[16] P.A. Lee and T. Anderson, "Fault tolerance: principles and practice," Springer-Verlag New York, Inc. Secaucus, 1990.

[17] W. Zhao, P.M. Melliar-Smith, and L.E. Moser, "Fault tolerance middleware for cloud computing," in IEEE 3rd International Conference on Cloud Computing (CLOUD 10), pp. 67-74, July 2010.

[18] P.T.T. Huyen and K. Ochimizu, "Toward inconsistency awareness in collaborative software development," in 18th Asia Pacific Software Engineering Conference (APSEC), pp. 154-162, Dec. 2011.

[19] S. Pallickara and G. Fox, “NaradaBrokering: a distributed middleware framework and architecture for enabling durable peer-to-peer grids,” *In* Proceedings of the ACM/IFIP/USENIX International Conference on Middleware (Middleware '03), pp. 41-61, 2003.

[20] S. Pallickara and G. Fox, "A scheme for reliable delivery of events in distributed middleware systems," in Proceedings of the IEEE International Conference on Autonomic Computing (ICAC'04), New York, NY, pp. 328-329, May 2004.

[21] S. Pallickara, M. Pierce, G. Fox, Y. Yan, and Y, Huang, "A Security framework for distributed brokering dystems", Available from http://www.naradabrokering.org, <retrieved: 03, 2012>.

[22] K. Komathy, V. Ramachandran, and P. Vivekanandan, “Security for XML messaging services: a component-based approach,” Journal of Network and Computer Applications, Vol. 26, Iss. 2, pp. 197-211, April 2003, DOI=10.1016/S1084-8045(03)00003-1 http://dx.doi.org/10.1016/S1084-8045(03)00003-1.

[23] F. T. Ammari and J. Lu, “Advanced XML security: framework for building secure XML management system (SXMS),” In Proceedings of the Seventh International Conference on Information Technology: New Generations (ITNG '10), Washington, DC, pp. 120-125, 2010, DOI=10.1109/ITNG.2010.124 http://dx.doi.org/10.1109/ITNG.2010.124.

[24] C. Narasimham and J. Pradhan,“Evaluation of performance characteristics of cryptosystem using text files”, Journal of Theoretical and Applied Information Technology,Vol. 4, Iss. 1, pp. 56-60, 2008.

Using CBR Method for Multi-Agent Selection of Multiple and Dynamic Composite Web Service

Fatma Siala
University of Tunis
SOIE - ISG
Tunis, Tunisia
fatma.siala@gnet.tn

Khaled Ghedira
University of Tunis
SOIE - ISG
Tunis, Tunisia
khaled.ghedira@isg.rnu.tn

***Abstract*—It is well-known that Web service technologies provide an easy way to integrate the applications within and across organizational boundaries. Web services are usually overlapping in functionality and how to make a choice based on non-functional factors becomes a problem that needs to be solved. This paper, argues that the selection of component services should be considered in a global manner based on the Web services availability and the users QoS preferences. Indeed, QoS becomes one of the most important factors for Web service selection. However, for a composition, we can have different combinations and execution paths. Particularly, a composite service can generate different schemes that give various QoS scores. This paper presents a framework which deals with the selection of composite Web services on the base of Multi-Agents negotiation and CBR (Case Based Reasoning) method. The objective of the agents is to find out the best Composite QoS (CQoS) based on Web services availability and elementary Web services QoS. By using CBR method, agents can memorize QoS scores. This framework supports different combinations and execution paths. The proposed Multi-Agents framework is compared to an existing approach in terms of execution time. Experiments have demonstrated that our framework provides reliable results in comparison with the existing one.**

***Keywords*-Web service; QoS; Multi-Agent System; Contract-Net Protocol; execution paths; availability; CBR technique.**

I. INTRODUCTION

Economical context impacts companies and their Information Systems (IS). Companies acquire other competitors or develop new business skills, delocalize whole or part of their organization. Their IS are faced to these complex evolutions and have to overcome these changes. In this context, Service Oriented Architecture (SOA) offers a great flexibility to IS. Applications are seen as black boxes independently connected to an application as Enterprise Application Integration bus (EAI) with its connectors. However, this integration solution does not allow connecting heterogeneous applications or infrastructures. Web services (WS) are based on standards and they are the cheaper and simplest solution to resolve this problem.

Service-Oriented Architecture (SOA) consists of a set of design principles which enable defining and composing interoperable services in a loosely coupled way. The value of SOA lies in assuring that such compositions are easily and rapidly possible with low costs. Thus, service composition is a key to SOA [22].

Yet, with the explosion of Web services available through out the Internet, it's not easy for the end users to composite the Web services manually to meet their specific preferences.

Quality of Web Service (QoS) has become a central criterion for differentiating competing service providers considering the increasing number of services with similar functionalities. The current service optimization paradigm assumes that precise QoS values are available for selecting the competing service providers [6], [22].

Moreover, a composite service can be represented by a statechart which has multiple execution paths when containing conditional branchings. Each execution path represents a sequence of tasks to complete a composite service execution. Furthermore, for a composite Web service, we notice that we can have different possible combinations.

Our work aims at advancing the current state of art in technologies for Web service composition by first, taking into account the user's preferences, second, by using agents to negotiate the execution path and combination offering the best QoS value, finally, by adding to agents the capability to memorize the QoS and the availability of each Web service.

By negotiating only with available Web services providers fulfilling the QoS user requirements, we obtain a better Central Processing Unit (CPU) time. The information about the QoS is memorized in a cases base. This framework improves the existing approaches [17] [16] in terms of CPU time when the agents can memorize the Web service availability and QoS.

Our contributions are revealed when negotiating only with available Web services providers and fulfilling the QoS requirements. We are able to memorize the QoS information by using CBR (Case Based Reasoning) method. This contribution gives a better Central Processing Unit (CPU) time and also supports different execution paths and combinations for a composition. The combination concept has never been addressed by any approach.

The structure of this paper is as follows: The coming section presents the major related works in the area of Web service composition based on QoS. Section 3 proposes the framework prototype based on Multi-Agent System and using CBR technique. Section 4 explains the implementation of this framework using a case study. Accordingly, Section 5 presents the experimentation results. After that, the approach is discussed in Section 6 by presenting existing approach limitations and detailing our contributions. Finally, we conclude the whole paper.

II. RELATED WORKS

Query optimization, taking into account QoS requirements, on Web services have received considerable attention in the service computing community [3] [14] [11]. In this section, we review selected works based on their relevance for our approach.

Zeng et al. [23] have presented a solution for the composition problem by analyzing multiple execution paths of a composite service which are specified using UML (Unified Modeling Language) statecharts. They have modeled the composition problem using different approaches, including a local optimization approach and global planning approach using linear programming.

Guan et al. [7] are the first who propose a framework for QoS-guided service compositions which uses constraint hierarchies as a formalism for specifying QoS. They use a branch and bound algorithm that is only capable of solving sequential compositions. The authors do not present any empirical evaluation to demonstrate the optimization performance of their approach.

Rosenberg et al. [15] have used constraint programming and integer programming approach for optimizing QoS by leveraging constraints hierarchies as a formalism to represent user constraints (specified with a Domain-Specific Language) of different importances.

Canfora et al. [4] have proposed an approach based on genetic algorithms. To determine the optimal set of concretizations, the approach needs to estimate the composite service QoS. This is done using some aggregation formulas.

Hong and Hu [8] have used an ordinary utility function as a numerical scale of ordering local services and a multi-dimension QoS based local service selection model is proposed to provide important grounds to choose a superior service and shift an inferior one. Secondly, subjective weight mode, objective weight mode, and subject-objective weight mode are constructed to determine the weight coefficient of each QoS criterion, and to show the users' partiality and the service quality's objectivity.

Alrifai and Risse [1] have employed Mixed Integer Programming (MIP) to find the optimal decomposition of global QoS constraints into local constraints. They have used distributed local selection to find the best Web services that satisfy these local constraints.

Yan et al. [20] have presented a framework in which the service consumer is represented by a set of agents who negotiate QoS constraints specified using SLA (Service Level Agreement), with the service providers for various services in the composition applying the Contract-Net protocol. Their idea of using Multi-Agents System and the Contract-Net protocol is in line with the work presented in this paper. However, the authors do not deal with the the Web dynamism (the availability concept) so their approach takes a large CPU time. Moreover, their framework does not support the different execution paths and combinations.

Most of these approaches does not take into account that for a composite service we can have an execution plan that generates different execution paths. These approaches deal only with the optimization problem itself (finding the best CQoS) without giving prominence for this aspect. Some approaches deal with this aspect but repeating each time the generation of all the elementary Web services. Moreover, all these approaches do not take into consideration that for a composite Web service we can have different combinations. This concept has never been addressed by any approach. By using Multi-Agents System, we generate all the execution paths and combinations in parallel and select the best execution path or combination, so we gain in terms of CPU time.

We also take into consideration the importance of the Web services' availability since the Web is a dynamic environment. By considering only available Web services we also improve the CPU time. This CPU time improvement is also due to using CBR method. In fact, agents memorize QoS scores for further user.

Our approach allows to find the CQoS that fulfill the user requirements by considering different execution paths and combinations and by also taking into account the dynamical aspect of the Web (the Web service availability and QoS scores changes). Our framework takes a largely better CPU time than the existing approaches.

III. THE PROPOSED FRAMEWORK

The first step toward autonomously establishing QoS value for a service composition is to have a supporting framework. This framework should be able to address the special requirements for establishing QoS value for a service composition. For the composition process, we use the technique described in [12] based on CBR method.

Our goal is to propose an approach to the Web services composition that guarantees non functional properties (QoS). This approach must also support different execution paths and combinations. Our framework allows to select the best elementary Web services in terms of QoS using CBR method and based on Web services availability and supports different execution paths and combinations.

Several motivations lead to use Multi-Agent Systems (MAS) [2]. In fact, a Web service suffers from 3 main deficiencies : it acknowledges only itself, passive until it is invoked, has only a knowledge of itself and neither adaptable nor able to benefit of new capabilities of the environment [5]. We propose to represent each service by an agent. In the MAS field, negotiation is a fundamental aspect of agents interactions. Indeed, since agents are autonomous, there is no imposed solution in advance, but agents must reach solutions dynamically while solving problems. The basic assumption of this approach is that each elementary Web service has an agent responsible to it. The objective of these agents is to find out the best CQoS. They recognize the available providers which are also represented with agents and the QoS associated to each of them using CBR technique.

We proposed in [17] a first framework, named Multi-Agent Availability (MAA), which optimizes the QoS criteria for a composite service provision and improves an existing approach [20] in terms of CPU time based on Web service availability. We have also proposed in [16] a second framework , named Multi-Agent Availability by exploring multiple execution Paths (MAAP) which improves the first one by considering different execution paths and combinations for a given statechart.

We compare our proposed frameworks with Yan et al.'s one [20]. To the best of our knowledge, their work is the sole which uses agent technology. Maamar et al. [13] have used MASs for Web service composition but they don't consider the QoS.

The MAA framework improves the work of [20] in terms of CPU time. In fact, we decrease the number of negotiations (we alleviate the network) by using a variant of the Contract-Net protocol, called the directed award and sending the CFP only to the available Web Service Agents. This contribution gains on CPU time.

On the other hand, the MAAP framework improves the work of [20] and the MAA framework in terms of QoS by supporting multiple execution paths and combinations for a composition. The MAAP framework takes more CPU time but we demonstrated that the difference of CPU time is negligible compared to the improvement of QoS. Supporting different combinations for a composition is a novel idea introduced by our work.

The MAA framework optimizes the QoS at the local level and verifies after that if it ensures the QoS at the global level (CQoS). On the other side, the MAAP framework, by considering different execution paths and combinations, has more chances to ensure the QoS at the global level.

In this context, we propose a new framework named Multi-Agent Availability by exploring multiple execution Paths based on QoS ($MAAP_Q$). Experiments are conducted to prove the effectiveness of our approach when the agents well also known the QoS (using CBR method) of each Web service with their availability.

A. $MAAP_Q$: An agent based Framework

To better explain our approach, we present in Fig 1 the framework ($MAAP_Q$). This framework consists of an Interface Agent (IA), a Combination Coordinator Agent (CCA) and a set of Negotiator Agents (NAs) that negotiate with a set of Web Service Agents (WSAs).

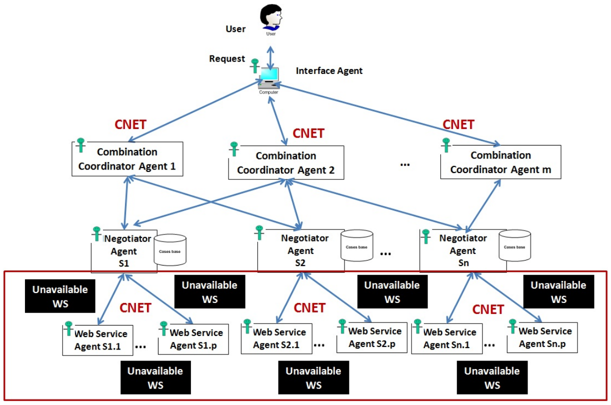


Figure 1. $MAAP_Q$'s Framework.

In the MAAP the negotiation process is as follows: First, the user specifies the desired composite Web service and the associated requirements. Then, The IA charges each CCA for an execution path or combination of the statechart. The CCA associates an NA for each elementary service in the composition. These NAs negotiate via the Contract-Net (CNET) protocol with WSAs (the providers) to find the best elementary Web service and send the response to the IA which evaluates the results and either confirms the acceptance or repeat the negotiation. Second, the IA negotiates via the CNET protocol for the best execution path. Finally, the IA returns to the user the best Web service composition.

We propose $MAAP_Q$ as an improvement of this framework when each NA posteriori knowns the WSA's QoS. So, we gain in terms of CPU time since the NA will not negotiate via the Contract-Net (CNET) protocol with WSAs (the providers) to find the best elementary Web service but prepare this information by a negotiation that takes place before this process. The NAs will only verify if this WSA fulfill the same QoS and respecting the user requirements and send the response to the IA which evaluates the results and either confirms the acceptance or repeat the negotiation.

We explain in the following the role of each agent.

1) The Interface Agent: The Interface Agent (IA) represents the interface that allows the user to access the framework for specify the desired service and QoS criteria. In fact, the user can specify each criterion with a weight since different users may have different requirements and preferences regarding QoS. For example, a user may require to minimize the execution time, while another user may give more importance to the price than to the execution time.

The selection process is based on the weight assigned by the user to each criterion. The IA computes the CQoS score using a formula proposed by [23]. The IA will then, provide the expected service with the required QoS criteria.

2) The Combination Coordinator Agent: The Combination Coordinator Agent (CCA) interacts with the IA to receive an execution path or combination of the statechart for the service composition and the user requirements. Each CCA is responsible for an execution path or combination of a composition. This agent attributes a NA for each Web services and computes the CQoS score using a formula proposed by [23] when it receives their responses. Finally, the CCA negotiates with the IA using the CNET protocol to provide its proposal.

3) The Negotiator Agent: A Negotiator Agent (NA) is responsible of a Web services set offering the same functionality. The NA uses CBR to have information about WSAs that check user's preferences.

If there is no case in the CBR that fulfill the user's requirements, the NA is in charge of negotiating with providers for a service from the composition in order to optimize the QoS. We use a variant of the CNET protocol, the directed award where the critical information which must be aware by the NA is the availability. So, the NA's acquaintances are the available Web service providers. For a negotiation, the NA must consult its list containing available Web services. This negotiation takes place before beginning the negotiation with the user. Each NA has its cases base for fulfillment with the QoS value associated with each QoS criterion.

4) The Web Service Agent: A Service Level Agreement (SLA) defines the terms and conditions of service quality that a Web service delivers to service requesters. The major constituent of an SLA is the QoS information. There are a number of criteria (e.g., execution time, availability) that contribute to a QoS in an SLA. Web service providers publish QoS information in SLAs.

Each Web Service Agent (WSA) represents a Web service. This Web service belongs to a Web services class where an NA is responsible. For example, a travel service provider may specify that it supports the Trip-planning service and belongs to the service class FlightTicketBooking. Service class describes the capabilities of Web services and how to access them.

The NA has a list that he consult whenever he begins a negotiation. In this list there is the available WSAs. By this method, we also take into account the failure cases. We note that in Fig 1, there are unavailable Web services represented by agents that will not negotiate with the NA.

B. How to apply CBR in $MAAP_Q$?

Using CBR by the NA consists of three steps the case representation, the case research and the case update.

1) Case representation: In case-based reasoning, an existing solution should have some similarity with the problem that is being solved in order to be reused. The $MAAP_Q$ considers properties associated to the available WSAs when matching it with existing instances of plans. These properties are the QoS' scores associated to each criterion of QoS (table I). Hence, to match with a Web service that fulfill a user requirements, an existing WSA must have the same or better value. Thus, checking that two WSAs are the same is to check whether they have the same QoS' scores. To make the matching of the WSAs efficient, a function defined in [21] is used to generate the digest of the BPEL file describing the workflow of a composite service. Thus, by comparing the digests of scores of two Web services, we can decide whether the Web service fulfill the user requirements. The NA will computes the QoS score based on the value associated to each criterion and the user preferences.

The instances of QoS scores stored in the CBR repository comply with Kolodner's work [10]. Each instance consists of three elements:

- Problem: A Web service;
- Solution: A selected WSA;
- Evaluation: the QoS score.

Table I
CASE REPRESENTATION.

	QoS_1	QoS_2	...	QoS_n
WSA_i	$value_1$	$value_2$	...	$value_n$

2) Case research: When a NA wants to select a WSA, it first checks the CBR repository to find out the WSAs that fulfill the user's requirements. The NA searches the request in the table I). If there are matching WSAs the NA computes its associated QoS fulfilling he user's preferences. The NA chooses the best provider (that fulfill the QoS requirements) and verifies its availability and if it fulfills the same values of QoS. In this case, The NA returns the result to the CCA. The NA considers 10 percent of the cases. Otherwise, if there are no cases that satisfy the user's request in the CBR, the NA negotiates with the WSAs via the contract-Net protocol.

3) Case update: After each negotiation between the NA and the WSAs, the NA updates the CBR with the new case for possible future use.

C. The Negotiation

1) Negotiation Protocol: The negotiation protocol is the way and manner the negotiating parties interact and exchange information. It includes the way in which the offers and messages are sent to opponents. There are various negotiation protocols available in the research community. In this paper, we propose to use the CNET Protocol, designed by [19], and especially a variant named directed award where the manager must have a table of acquaintances that contains knowledge about other agents (eg, skills, knowledge, value judgments about these agents). In this protocol, one agent (the initiator) takes the role of manager which wishes to have a task performed by the other agents (the participants) and further wishes to optimize a function that characterizes this task. We use the function proposed by [23]. For a given task, any number of participants may respond with a proposal, the rest (agents that cannot respect QOS requirements) must refuse.

2) Negotiation Coordination: The IA coordinates multiple negotiations for various services in the composition. The purpose of the coordination is to ensure that the results of these multiple negotiation can collectively fulfill the end-to-end QoS. The first task that the IA performs is to attribute a combination or an execution path of a composite Web service with its QoS weights to the CCAs. The second task that the IA performs is to confirm negotiation results. As the extended negotiation protocol suggests, when various CCAs get the best deals, they consult the IA for confirmation. The IA evaluates the results and either confirms the acceptance or amends the reserve values to continue the negotiation. The QoS aggregation refers to the QoS model proposed in [23]. Each CCA charges the NAs to find the best elementary Web services.

If the NAa does not find in the cases base WSAs that fulfill the user requirements, it sends a CFP message only to the available Web Service Agents. When The proposals and counter proposals are then communicated iteratively between the NAs and the service providers (WSAs), following the standard FIPA protocol, until the best deal is reached (i.e. the proposals offered by one or more providers can satisfy the negotiation objectives) or the timeout occurs. The NA blocks the selected WSA. At this time, the best deal is sent to the IA. If the overall QoS requirements are satisfied, the IA confirms to each CCA that the current deal is acceptable. Subsequently, the CCA acknowledges the acceptance of the proposal to the NAs that must inform the selected providers (WSAs). When the user finishes using Web services, the IA informs the NA to unlock the selected WSA.

In case the overall QoS requirements are not satisfied based on the current best deals, the user should modify the requirements and the IA amends the reserve values for the CCA to re-start negotiation. Each corresponding NA then sends the modified CFP to WSAs and begins a new negotiation process

This research refers to the QoS model presented in [23] which proposes a formula to compute the overall QoS score for each Web service.

For a given task, the NA will choose the Web service which satisfies all the user preferences for that task and which has the maximal score. If there are several services with maximal score, one of them is selected randomly. If no service satisfies the user preferences for a given task, an execution exception will be raised and the IA will propose the user to change his preferences.

IV. IMPLEMENTATION AND CASE STUDY

To show the key ideas presented in this paper, a prototype has been implemented for the proof-of-concept purpose using the FIPA compliant JADE (Java Agent Development Framework) [9] which is a middleware that implements an agent platform and a development framework. This framework supports CBR and agents' negotiations through an Agent Communication Language (ACL).

During the negotiation, the IA, the CCAs, the NAs and the WSAs exchange a number of messages.

We explain our approach through a case study. A simplified statechart specifying a scenario in the tourism industry composite Web service is depicted in Fig 2.

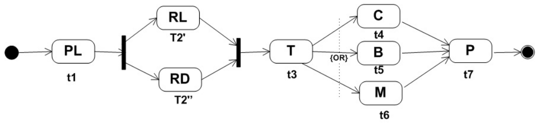


Figure 2. Example of a statechart.

In this scenario, a tourist who holds a mobile device can request the full description of the route information from his/her current position to a selected attraction. We have height different services, that will be invoked. A Phone Location Service (S_{PL}), a Route Calculation Service (S_{RL}), a Route Description Service (S_{RD}), a Traffic Service (S_T), a Car Service (S_C), a Bus Service (S_B), a Metro Service (S_M) and a Metro Service (S_P). The tourist can also specify some QoS requirements when making his/her request. For example, the tourist can request that the score of CQoS is delivered with a value above 70. Obviously, the tourist can also require the QoS score for the elementary Web services such S_V, S_B, S_M, etc.

The tourist should also indicate the weight associated to each QoS attribute. These weights will be used for the computation of the QoS score of each elementary Web services using a formula proposed in [23]. If the user does not specify weights, the system will consider a weight value = 0.25 for each criterion.

For a user's request, each IA sends at the same time a CFP to the CCAs for an execution path or combination in the statechart of the composition. In our case, we have three execution paths. Each NA (height NAs) associated to an elementary Web service checks the QoS' user requirements with the WSAs existing in its cases base. For example, if the user specifies the following weights values : price=0.15, duration=0.35, success rate=0.40 and reputation=0.10. The NA will apply the formula proposed by [23] to choose the best WSA. This formula is based on the QoS scores and their associated weights. Table II represents an example of different values offered by various WSAs associated to S_{PL}.

After that, each NA verifies if these providers are again available and fulfill the user requirements in term of QoS.

If this verification is true, the NA returns the result to the CCA. The NA considers 10 percent of the cases. Else, (if there is no cases that satisfy the user's request in the CBR or the WSAs are not available), the NA negotiates with the WSA via the Contract Net protocol. In this case we use the MAAP process [16].

Table II
EXAMPLE OF CASE REPRESENTATION.

WSA, QoS	*price*	*duration*	*success*	*reputation*
S_{PL1}	40	55	70	80
S_{PL2}	60	65	69	73
S_{PL3}	70	68	64	71

Negotiation is as follows: Each NA communicates with providers simultaneously. The negotiation results for each service are summarized in Table III. In all cases, after the reception of offers from available WSAs, the height NAs select the best offer, block the WSAs associated to the selected Web services and return their best offers to the CCAs. Each CCA (three CCAs) calculates its CQoS score. Supposing that we have these results after the negotiation, EP1= 65; EP2 = 84 and EP3=40. The IA will choose the EP2 that has the best CQoS. The whole process is comprehensively simulated using the prototype.
The following results are confirmed: S_{PL}: 66, S_{RL} : 98, S_{RD} : 80, S_T : 79, S_C:87, S_B : 73, S_M : 85 and S_P : 92.

These results are associated to the best QoS of each elementary Web service. Finally, the IA checks if the best offers can jointly fulfill the user's request (the desired value of CQoS is 75) using a formula also proposed by [23].

When the tourist finishes using the Web service, the IA informs the NA to unlock the selected (reserved) WSA.

Table III
QOS VALUES (NEGOTIATION RESULTS) FOR EACH WSA.

	S_{PL}	S_{RL}	S_{RD}	S_T	S_C	S_B	S_M	S_P
1	65	46	80	25	35	73	85	92
2	33	39	40	48	28	54	77	45
3	66	24	50	38	66	49	76	22
4	40	45	26	79	32	46	80	21
5	22	98	44	75	87	24	37	64

V. EXPERIMENTATION

The series of tests were conducted to compare the CPU time of $MAAP_Q$ framework with MAA.

In the experimentation, we have calculated the CPU time for each approach by varying the number of elementary services in a composition from 5 to 50 with steps of 5 and varying the number of service providers from 10 to 50 with steps of 10. We have calculated the CPU time 10 times for each case and we have considered the average. We present via a 3D chart in Figure 3 the results of these experiments. The results of *MAA* are represented in blue and the result of $MAAP_Q$ approach are represented in green. These experiments show that the $MAAP_Q$ takes a far better result than *MAA*. For example, if the number of Web services providers is equal to 50 and the number of elementary services in a composition is equal to 30, the $MAAP_Q$ takes 2300 ms but *MAA* takes 3035 ms. We gain in terms of CPU time.

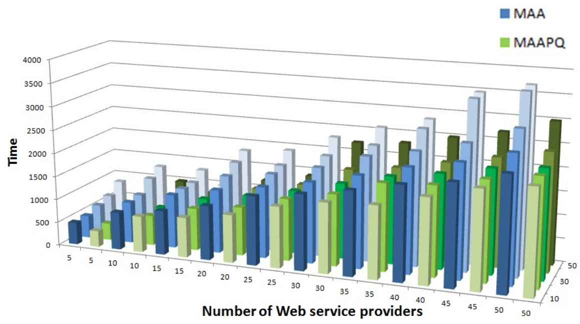


Figure 3. Comparison of CPU time.

We show through Figure 4 the gain percentage of the $MAAP_Q$ compared to the MAAP framework (using formula 1).

$$Gain = \frac{CPU_{time}(MAA) - CPU_{time}(MAAP_Q)}{CPU_{time}(MAA)} * 100 \qquad (1)$$

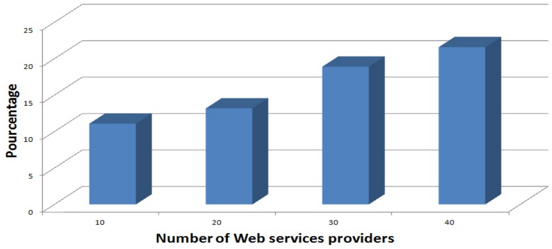


Figure 4. Gain Percentage.

VI. DISCUSSION

The MAA framework optimizes the QoS at the local level and verify after that if it ensures the QoS at the global level (CQoS). On the other side, the MAAP framework, by considering different execution paths and combinations, has more chances to ensure the QoS at the global level. By using CBR method, agents memorize QoS scores for further user. this technique allows us to gain in term of CPU time. It is a method to limit the research space.

By using the CBR method, NAs have knowledge about the WSAs. So, we gain the time of conversation. The $MAAP_Q$ not only offers a QoS that fulfill the user requirements but also takes a largely lower CPU time than these frameworks.

Note that our approach supports as more attributes of QoS (price, duration, reputation, success rate, availability, etc.) as we want.

VII. CONCLUSION

This paper exploited the agent technology to select composite Web service using CBR method. This technique allows agents to memorize the QoS scores for further use. The system reduces the amount of time spent on solving a service composition problem by reusing previous selected Web services. The experiment shows that, when there are sufficient amount of Web services in the CBR repository, the proposed system can outperform the existing approaches.

Our approach advances the current state of the art by taking into account the Web services availability and supporting different execution paths and combinations for a composition. By using CBR method we gain in terms of CPU time.

The greatest limitation of this framework is its lack of scalability. Therefore, we present in [18] a new scalable framework using Case Based Reasoning.

In the future work, we will propose a new approach which improves again the execution time by discharging the NAs of many tasks and by adding horizontal communications.

REFERENCES

[1] Mohammad Alrifai and Thomas Risse. Combining global optimization with local selection for efficient qos-aware service composition. In *WWW*, pages 881–890, 2009.

[2] Mihai Barbuceanu and Mark S. Fox. The design of a coordination language for multi-agent systems. In *ATAL*, pages 341–355, 1996.

[3] Ivona Brandic, Sabri Pllana, and Siegfried Benkner. Specification, planning, and execution of qos-aware grid workflows within the amadeus environment. *Concurrency and Computation: Practice and Experience*, 20(4):331–345, 2008.

[4] Gerardo Canfora, Massimiliano Di Penta, Raffaele Esposito, and Maria Luisa Villani. A framework for qos-aware binding and re-binding of composite web services. *Journal of Systems and Software*, 81(10):1754–1769, 2008.

[5] Yasmine Charif, Kostas Stathis, and Hafedh Mili. Towards anticipatory service composition in ambient intelligence. In *NOTERE*, pages 49–56, 2010.

[6] Francisco Curbera, Bernd J. Krämer, and Mike P. Papazoglou, editors. *Service Oriented Computing (SOC), 15.-18. November 2005*, volume 05462 of *Dagstuhl Seminar Proceedings*. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2006.

[7] Ying Guan, Aditya K. Ghose, and Zheng Lu. Using constraint hierarchies to support qos-guided service composition. In *ICWS*, pages 743–752, 2006.

[8] Liurong Hong and Jianqiang Hu. A multi-dimension qos based local service selection model for service composition. *JNW*, 4(5):351–358, 2009.

[9] JADE. Telecom italia lab. In *http://sharon.cselt.it/projects/jade*, 2011.

[10] J. L. Kolodner. Case-based reasoning. In *Morgan Kaufman*, 1993.

[11] Srividya Kona, Ajay Bansal, M. Brian Blake, Steffen Bleul, and Thomas Weise. Wsc-2009: A quality of service-oriented web services challenge. In *CEC*, pages 487–490, 2009.

[12] Soufiene Lajmi, Chirine Ghedira, Khaled Ghédira, and Djamal Benslimane. Wescocbr: How to compose web services via case based reasoning. In *ICEBE*, pages 618–622, 2006.

[13] Zakaria Maamar, Soraya Kouadri Mostéfaoui, and Hamdi Yahyaoui. Toward an agent-based and context-oriented approach for web services composition - appendices. *IEEE Trans. Knowl. Data Eng.*, 17(5), 2005.

[14] Arun Mukhija, Andrew Dingwall-Smith, and David S. Rosenblum. Qos-aware service composition in dino. In *ECOWS*, pages 3–12, 2007.

[15] Florian Rosenberg, Predrag Celikovic, Anton Michlmayr, Philipp Leitner, and Schahram Dustdar. An end-to-end approach for qos-aware service composition. In *EDOC*, pages 151–160, 2009.

[16] Fatma Siala and Khaled Ghédira. A multi-agent selection of multiple composite web services driven by qos. In *CLOSER*, pages 675–684, 2011.

[17] Fatma Siala and Khaled Ghédira. A multi-agent selection of web service providers driven by composite qos. In *ISCC*, pages 55–60, 2011.

[18] Fatma Siala, Soufiene Lajmi, and Khaled Ghédira. Multi-agent selection of multiple composite web services based on cbr method and driven by qos. In *iiWAS*, pages 90–97, 2011.

[19] Reid G. Smith. The contract net protocol: High-level communication and control in a distributed problem solver. *IEEE Trans. Computers*, 29(12):1104–1113, 1980.

[20] Jun Yan, Ryszard Kowalczyk, Jian Lin, Mohan Baruwal Chhetri, SukKeong Goh, and Jian Ying Zhang. Autonomous service level agreement negotiation for service composition provision. *Future Generation Comp. Syst.*, 23(6):748–759, 2007.

[21] Xinfeng Ye and Rami Mounla. A hybrid approach to qos-aware service composition. In *ICWS*, pages 62–69, 2008.

[22] Tao Yu, Yue Zhang 0001, and Kwei-Jay Lin. Efficient algorithms for web services selection with end-to-end qos constraints. *TWEB*, 1(1), 2007.

[23] Liangzhao Zeng, Boualem Benatallah, Anne H. H. Ngu, Marlon Dumas, Jayant Kalagnanam, and Henry Chang. Qos-aware middleware for web services composition. *IEEE Trans. Software Eng.*, 30(5):311–327, 2004.

Hybrid Recommendation of Composition Knowledge for End User Development of Mashups

Carsten Radeck, Alexander Lorz, Gregor Blichmann, Klaus Meißner
Technische Universität Dresden, Germany
{*Carsten.Radeck,Alexander.Lorz,Gregor.Blichmann,Klaus.Meissner*}*@tu-dresden.de*

***Abstract*—Due to the increasing number of available web APIs and services, mashups have become a prominent approach for building situational web applications. Still, current mashup tooling is not suitable for end users lacking detailed understanding of technical terms or development knowledge. A promising approach to ease mashup development is end user guidance utilizing recommendations on composition knowledge gained from experienced, similar users and semantic component annotations. In this paper, we present a novel hybrid recommendation approach suggesting suitable advice for the application development and adaptation to the end user. Composition knowledge in terms of common composition patterns is applied. Pattern instances are generalized by determining semantic exchangeablity of components to allow for context-sensitive recommendations. In addition, they are automatically integrated with the running application.**

***Keywords*-Mashup; End User Development; Hybrid Recommendation; Runtime Composition.**

I. Introduction

While the amount of available application programming interfaces (APIs) and third party resources in the Web is steadily increasing, the emerging *mashup* paradigm enables loosely coupled application components to be reused in several scenarios by simply combining them. Besides conventional mashup tools focussing on aggregation and processing of data from heterogeneous sources, there are proposals for the *universal composition* of mashups such as [1]. It includes the uniform composition and integration of distributed web resources. The latter are encapsulated in uniformly described components spanning all application layers including the user interface.

End User Development (EUD) aims at supporting the construction of individual applications addressing situational user needs. To this end, mashups are in principle a promising approach due to the reuse of building blocks and lower development efforts. We address domain experts, not necessarily with programming or technical knowledge, as end users. For those, early mashup platforms, e. g., Yahoo! Pipes, are unsuitable since they usually require an understanding of underlying technologies or programming skills. Furthermore, given the increasing amount of web resources and services, finding the right components or compositions is a challenging task, especially if there are only technical interface descriptions like WSDL. To overcome this and to foster EUD, recommendations for meaningful composition steps are of increasing importance [2], and gain momentum in the mashups domain as well. Our approach of universal composition by the given target group of end users at the application's runtime leads to several *challenges*:

C1 Besides the classic cold start problem, taking context-sensitivity and QoS into account is a main challenge in the domain of web services [3]. Given universal composition, this is of particular importance since we apply binding and instantiation of components in the runtime environment and support different target platforms. Compositions should also be restorable on different platforms later on. In such a setting, the components' suitability to the current context, especially device and software capabilities, has to be guaranteed.

C2 Since mashups claim to fulfil long-tail user needs, not only the "most popular" components and composition steps should be considered.

C3 The reason for and origin of recommendations should be presented to the end user [4]. Additionally, we argue that recommending functionality based on composition knowledge should be applied to hide technical details.

C4 The ad-hoc reconfiguration of the mashup at runtime necessitates the actual integration of recommended composition parts and should be highly automated. In addition, interface heterogeneity of semantically compatible components should be resolved automatically.

We propose a novel hybrid recommendation approach presenting suitable advice for composition steps to the end user. Therefore, we leverage composition knowledge in terms of patterns, either mined statically from existing applications or dynamically based on the lightweight semantic component annotations. By reasoning on composition patterns' functionality, determining semantic exchangeablity of components and integrating patterns in running applications, we aim to overcome the limitations of prevalent solutions.

The remaining paper is structured as follows. Related approaches for recommending components or composition knowledge and the conceptual foundation of our work are briefly discussed in Section II. Next, our hybrid recommendation approach is presented in Section III. Section IV, finally, summarizes the vision and outlines work in progress.

II. Related Work and Foundation

In the following, we discuss related approaches and define our conceptual basis in more detail.

A. Related Work

In general, recommender systems can be classified as collaborative filtering, content-based, and hybrid [5]. While collaborative filtering approaches utilize preferences of users similar to the active user, content-based ones recommend items similar to the active user's preferences.

There are a number of approaches using semantic similarity of component interfaces, e. g., [6]. Such matching approaches work well for the determination of alternative and possible follow-up components including their coupling. Additionally, rather "uncommon" solutions are recommendable and cold start can be avoided. On the other hand, semantic component descriptors are required, which have to be sufficiently expressive and, thus, complex to calculate whole compositions. However, the proposals mentioned above do not consider context information, collective knowledge, and the integration of recommendations.

Other approaches build up on previously defined compositions of other similar users. [7] utilizes collective knowledge by reuse of *mashlets* and *glue patterns* to suggest missing components and connections. [8] extends this approach by ranking compositions with regard to multiple QoS and context criteria. Based on semantic component descriptors, semantic matching and AI planning, *MashupAdvisor* calculates compositions probably fitting user goals (desired outputs of the mashup) [9]. Thereby, the statistical (co-)occurrence of input and output concepts are derived from existing compositions. In wisdom-aware computing [10], composition knowledge is provided as *advices* associated with *patterns* comprising the actual knowledge; *triggers* state the condition under which the *advices* are offered. The proposal relies on implicit semantics gathered by different mining techniques and statistical data analysis on existing compositions. This work covers the integration of recommendations in mashups. These approaches suffer from cold start, prefer popular solutions, and lack awareness of device or user context.

Other work focusses on the derivation of composition recommendations, which can, for instance, be achieved by mining frequent web service sequences [11] or matching the *composition context* (a composition fragment of connected services around a certain service) [12]. However, they suffer from cold start and context-aware substitution of services or their integration with the mashup are not addressed.

Hybrid recommender systems for web services are proposed in [13] and [3]. Although they overcome cold starts, both recommend single web services only and dynamic service integration is not considered.

In summary, none of the solutions fulfils the requirements identified in Section I. The next sections present the prerequisites and the overall concepts of our envisioned approach.

B. Universal Composition in CRUISe

Universal composition is in our case provided by CRUISe, which allows for platform and technology independent composition of arbitrary web resources and services [1]. Components encapsulating these resources are uniformly described using the Semantic Mashup Component Description Language (SMCDL) [14]. SMCDL covers non-functional properties and the public component interface including functional and data semantics of operations, events, and properties by means of ontology concepts. A declarative composition model describes the mashup application including the components, their state, event-based communication, and layout [1]. *Templates* as part of composition models allow for the context-aware selection of semantically compatible components suitable for the target runtime environment and user preferences. To this end, templates are equally characterized by a component interface, but additionally include non-functional requirements for ranking candidates.

III. Recommendation approach

In this section, we outline our novel concept for an EUD mashup platform utilizing hybrid recommendation. We extend the basic concepts of related proposals in several directions due to the challenges (*C1–C4*) sketched above.

The overall approach is shown in Figure 1 and has similarities to an adaptation loop, which, separated from the application, includes continuous monitoring of the context and the application, analyzing (i. e., evaluation of trigger conditions and calculating recommendations in terms of patterns), planing (i. e., deriving an action specification), and adapting the application by executing the plan (i. e., realization of the action specification). Details on the main steps and concepts are provided next.

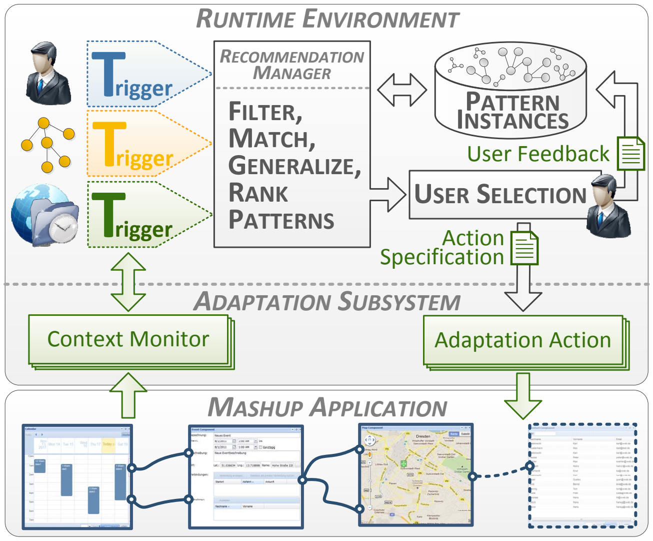


Figure 1. Overview of the recommendation approach

A. Triggering the Recommendation Process

We propose a unified trigger concept to cover the whole composition process starting from an empty application canvas to the interactive development at runtime. As an extension of [10], we distinguish *explicit* and *implicit* triggers. The former, colored blue in Figure 1, require the user's intention to ask for recommendations. The latter can be distinguished in *reactive* (colored green) and *proactive* (colored orange) and are generated by the platform based on context information. Reactive triggers respond to context changes, while proactive ones require the platform to continuously request and analyze additional information that might lead to recommendations. The following list provides examples:

- The user enters a brief description of his / her current task to be solved in a text field (*explicit*).
- Compared to compositions of similar users, a certain component is not part of the mashup (*proactive*).
- An event fired by a map component is not yet connected to any component (*reactive*).
- Entering a meeting room, a new service offered by a digital whiteboard becomes available (*reactive*).
- The underlying web service of a component has been unavailable for several requests (*reactive*).

Triggers model conditions setting them off, e. g., a user action and an affected component in case of an explicit trigger. Further, triggers are associated with a defined set of pattern classes, which are outlined in the next section.

Implicit, reactive triggers continuously receive notification of context changes. Those notifications are generated by the runtime environment's *adaptation subsystem* [15] that monitors context, like the user's position or device state. The *Recommendation Manager* interprets trigger events to suggest helpful composition fragments. As further data sources, it has access to the current composition model, a repository of composition models, and the user's context model.

B. Modelling and Querying the Composition Knowledge

Composition knowledge is represented by patterns describing common composition fragments, e. g., components and their connections. Pattern classes and pattern instances can by distinguished. In addition to the pattern classes *co-occurence*, *coupling*, *configuration*, and *complex* identified by [10], we introduce *occlusion*, *exchangeability*, and *layout*. While *exchangeable* components provide the same functionality, an *occluding* component offers additional functionality. *Layout* represents a typical arrangement of components.

Besides a characteristic composition fragment described by the composition model mentioned in Section II-B, each pattern has a rating and an origin to create trust and enable traceability. The origin describes whether the pattern was detected by semantic reasoning or from collective knowledge. A further attribute is the functionality, which can be derived from the components' semantic annotations. This allows for a user-appropriate visualization of the recommendations primarily showing *what* will happen and, secondarily, which composition fragments will be involved.

Pattern instances are decoupled from their mode of detection. Annotation-based semantic reasoning on exchangeable and connectible components takes place either statically or upon request. This way, the cold start problem is avoided and "niche requirements" can be met (*C2*). On the other hand, statistical analysis and data mining of existing compositions, for example, as proposed by [13] and [3], allow for utilization of (complex) collective composition knowledge.

Based on a trigger, a preselection of suitable patterns utilizing the mapping from trigger to pattern classes is conducted. Then, pattern instances are matched against the current composition and the conditions modelled by a trigger. To allow for context-sensitive recommendations (*C1*), we abstract pattern instances using *templates*.

Example: A mashup contains the component "Calendar", which is, as identified from collective knowledge, frequently connected with "Facebook Contacts" (*coupling*). The context model states that the user has no Facebook account, but is registered at Google+. A second component, "Google Contactor", is recognized as a semantic substitute for "Facebook Contacts" in relation to the coupling with "Calendar". Therefore, the coupling with "Google Contactor" is recommended.

Those equivalence classes of exchangeable components can be determined by semantic matching of component interfaces [14]. This way, we are able to generalize pattern instances by identifying the occurrence of abstract component classes, represented by templates, instead of concrete components. The selection of a concrete component for a template utilizes context information and non-functional annotations. Thus, the suitability of components for satisfying the user's preferences and their compatibility with a certain runtime platform and device are considered.

Collaborative filtering takes place to rank the patterns with respect to their popularity, ratings and relevance for the current user. Existing solutions can be leveraged for this task, e. g., clustering of users as proposed by [3].

Finally, the declarative *action specification* is derived. It represents the changes in the composition model and, thus, states all necessary steps to integrate a particular pattern instance with the current mashup.

C. Presenting and Integrating the Composition Knowledge

When visualizing suitable patterns for a trigger, the additional functionality provided to the user is emphasized (*C3*).

Example: The user requests the extension of the mashup containing the "Calendar". The functionality *Invite persons to your appointment* is offered. After the user agrees, a selection of appropriate pattern instances, containing amongst others the coupling with "Google Contactor", is shown.

One important requirement is the dynamic integration of the recommended composition knowledge with the mashup under development (*C4*). For this purpose, the defined action specification for the pattern instance selected by the user is the main input and interpreted by the runtime environment. Again, we utilize and extend the adaptation subsystem of the runtime environment since it provides the necessary means to realize adaptations at the component and the composition layer, cf. Figure 1. Amongst others, declarative *adaptation actions* for adding, exchanging, and reconfiguring components, adapting the layout, as well as creating communication channels, and registering publishers and subscribers are supported. This allows for a seamless integration of our concepts with the underlying infrastructure. The correct order of adaptation actions and their transaction-oriented execution are guaranteed by the adaptation subsystem.

Example: The action specification for the coupling pattern from the previous example comprises *adding "Google Contactor"*, *creating a communication channel between "Calendar" and "Google Contactor"*, and *registering the components as publisher or subscriber with this channel*.

The user's feedback on a pattern instance's suitability is collected, both explicitly by ratings and implicitly depending on whether the pattern instance has been applied or not.

IV. Conclusion and Future Work

In this paper, we have given a brief overview of our approach to recommend composition knowledge. The latter is derived from existing compositions of similar users (collaborative filtering) or from semantic component descriptions (content-based) and weaved it into running mashups. Based on a unified notion of recommendation triggers and composition patterns as well as an adaptation-enabled runtime, we provide continuous development support for end users during the usage of a mashup application. To allow for context-sensitive recommendations, patterns are generalized by semantically reasoned exchangeability of components.

The envisioned concepts are still early work in progress requiring further elaboration, e. g., detection mechanisms for implicit triggers as well as aggregation and processing of trigger events. We continuously work on prototypes within the CRUISe architecture to study practicability and feasibility in different application domains. One of the next steps is to incorporate and extend means for mediation resulting from our previous work [14]. Further, a user study is planned to evaluate the recommendation approach.

V. Acknowledgments

Funding for the EDYRA project is provided by the Free State of Saxony and the European Union within the European Social Funds program (ESF-080951805).

References

[1] S. Pietschmann, V. Tietz, J. Reimann, C. Liebing, M. Pohle, and K. Meißner, "A metamodel for context-aware component-based mashup applications," in *12th Intl. Conf. on Information Integration and Web-based Applications & Services (iiWAS 2010)*. ACM, Nov. 2010, pp. 413–420.

[2] A. Namoun, T. Nestler, and A. De Angeli, "Service composition for non-programmers: Prospects, problems, and design recommendations," in *8th European Conference on Web Services (ECOWS 2010)*. IEEE, Dec. 2010, pp. 123–130.

[3] L. Liu, F. Lecue, and N. Mehandjiev, "A hybrid approach to recommending semantic software services," in *Intl. Conf. on Web Services (ICWS 2011)*. IEEE, Jul. 2011, pp. 379–386.

[4] A. De Angeli, A. Battocchi, S. Roy Chowdhury, C. Rodriguez, F. Daniel, and F. Casati, "End-user requirements for wisdom-aware eud," in *End-User Development*, ser. LNCS. Springer, 2011, pp. 245–250.

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

[6] D. Bianchini, V. De Antonellis, and M. Melchiori, "A recommendation system for semantic mashup design," in *Workshop on Database and Expert Systems Applications (DEXA)*, 2010, pp. 159–163.

[7] O. Greenshpan, T. Milo, and N. Polyzotis, "Autocompletion for mashups," *Proc. of the VLDB Endowment*, vol. 2, no. 1, pp. 538–549, Aug. 2009.

[8] M. Picozzi, M. Rodolfi, C. Cappiello, and M. Matera, "Quality-based recommendations for mashup composition," in *Current Trends in Web Engineering*, ser. LNCS. Springer, Jul. 2010, pp. 360–371.

[9] H. Elmeleegy, A. Ivan, R. Akkiraju, and R. Goodwin, "Mashup advisor: A recommendation tool for mashup development," in *International Conference on Web Services (ICWS 2008)*. IEEE, Sep. 2008, pp. 337–344.

[10] S. Roy Chowdhury, C. Rodríguez, F. Daniel, and F. Casati, "Wisdom-aware computing: On the interactive recommendation of composition knowledge," in *Service-Oriented Computing*, ser. LNCS. Springer, Dec. 2010, pp. 144–155.

[11] A. Maaradji, H. Hacid, R. Skraba, and A. Vakali, "Social web mashups full completion via frequent sequence mining," in *World Congress on Services (SERVICES 2011)*. IEEE, Jul. 2011, pp. 9–16.

[12] N. Chan, W. Gaaloul, and S. Tata, "Composition context matching for web service recommendation," in *Intl. Conf. on Services Computing (SCC 2011)*. IEEE, 2011, pp. 624–631.

[13] C. Zhao, C. Ma, J. Zhang, J. Zhang, L. Yi, and X. Mao, "Hyperservice: Linking and exploring services on the web," in *Intl. Conf. on Web Services (ICWS 2010)*. IEEE, Jul. 2010, pp. 17–24.

[14] S. Pietschmann, C. Radeck, and K. Meißner, "Semantics-based discovery, selection and mediation for presentation-oriented mashups," in *5th Intl. Workshop on Web APIs and Service Mashups (Mashups)*. ACM, Sep. 2011, pp. 1–8.

[15] S. Pietschmann, C. Radeck, and K. Meißner, "Facilitating context-awareness in composite mashup applications," in *3rd Intl. Conf. on Adaptive and Self-Adaptive Systems and Applications (ADAPTIVE 2011)*. XPS, Sep. 2011, pp. 1–8.

DLConnector: Connecting a publication list to scholarly digital library

Jen-Ming Chung[1], William W.Y. Hsu[1], Cheng-Yu Lu[1*], Kuo-Ping Wu[2], Hahn-Ming Lee[2], and Jan-Ming Ho[1]
[1]*Institute of Information Science, Academia Sinica, Taiwan*
{*jenming,wwyhsu,cylu,hoho*}*@iis.sinica.edu.tw*
[2]*Dep. CSIE, National Taiwan University of Science and Technology, Taiwan*
{*wgb,hmlee*}*@mail.ntust.edu.tw*
[*]*corresponding author*

***Abstract*—Researchers and graduate students spend a great deal of time on searching and reading papers. However, they may usually lack of experience on using proper keywords and finding the relevant papers can be challenging. To keep track on the newest activities of key authors and venues is important as well. In addition, some researcher present publication list on his/her home pages. To maintain the up-to-date information of the publication list for the web pages (i.e., citation number) is labor-intensive. In this paper, we incorporate our previous work and construct a Web 2.0 platform, namely *Digital Library Connector* (*DLC*). *DLC* provides services to facilitate the tedious research processes. Researchers can easily search relevant papers and subscribe other author's academic activity. Moreover, researchers can easily construct their Web 2.0 web pages to present their profile, publication list, and recent academic activity by using the service on *DLC*. The users show highly satisfactory feedback on using *DLC*.**

Keywords-Key Paper Searching; Key Author Recommendation; Expertise Extraction; Web 2.0 Scholarly Platform;

I. Introduction

A serious literature survey of a new research topic for a beginning graduate student is an essential work. However, students may have limited experience on his/her research topic, and usually spend much time on searching and reading academic papers. This process, namely *Research Cycle* [21], is described as follow: *Using keywords to digital library*, *Reading papers from the returned results*, *Focus on key papers*, *Keeping track on important authors and venues*, and *Focus on research society*. One of the scenarios is that a student submits a list of queries to scholarly search engines to retrieve relevant papers. However, using proper keywords can be challenging, especially for beginners who are lack of experience and domain knowledge. To find relevant papers from huge returned results is time-consuming which may lead to much wasted effort to make determination of importance of the papers. Furthermore, it is necessary to keep track of the latest research, activities from the academic society.

In addition, experienced scholars usually need to maintain the self-owned publications for sharing the latest achievements and describing their research contributions. The most common way is to make a home-page or use the template provided by their institute. These web pages may usually not provide academic information for audience, i.e., citation number that partly represents the impact of the research topic.

In this paper, we incorporate with our recent [1], [2], [3], [4], [5] and some of previous work [6], [7], [8], [9], [10], [11] to develop a platform for scholarly use, namely *Digital Library Connector* (*DLC*), to facilitate the process (research cycle). The *DLC* provides researchers services as follows: (*i*) *Key paper searching*: searching the publishing lecture about the target conceptual entity; (*ii*) *Key author recommendation*: recommending the authors with the closely relevance with the given topic; (*iii*) *Expertise extraction*: extracting expertise by analyzing publications; (*iv*) *Web 2.0 platform construction*: providing services for researchers sharing and maintaining the publication data, users subscribe researchers and papers. The remainder of this paper is organized as follows. In Section 2, we review the state-of-the-art research, related work, and our previous research results. We first formulate a topic model for multiscale dynamics, and describe its online inference procedures. In Section 3, the platform *Digital Library Connector* is proposed, we elaborate the system architecture and components. In Section 4, we demonstrate the effectiveness of the proposed method by analyzing the dynamics of real document collections. Finally, discussion and conclusion are presented.

II. Related Work

To develop an academic platform to support the research cycle requires to keep track on (1) Relevant papers retrieval. Researchers need focus on not only the newest papers but also the classic papers of a topic (2) Key authors and venues localization; Key authors usually present advanced research on famous conference/journals (3) A platform that provides these services.

A. Key Paper Searching

Chen *et al*. [4] proposed a citation-network-based methodology, namely Citation Authority Diffusion (CAD), to rapidly mine the limited key papers of a topic, and measure the novelty on literature survey. A defined Authority Matrix (AM) is used to standardize duplication rate of authors and to describe the authority relation between the citing and

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

the cited papers. Based on AM, our CAD methodology leverages the Belief Propagation to diffuse the authority among the citation network. Therefore, CAD transforms the citation network to a novelty paper list to researchers. The experimental results show CAD can mine more novelty papers by using real-world cases.

B. Key Author Localization

Wu *et al*. [3] presented an approach applying Web Mining to recommend key authors of a topic. The authors designed a measure, namely *p*-index, for the ranking of researchers. Users can use academic keywords such as ”Data Mining”, the service returned a list of ranked authors.

C. Expertise Extraction

Lu *et al*. [2] and Yang *et al*. [8] analyzed researcher's publication list and turned out their expertise. They both adopt *Wikipedia* as ontology which contains many fashion terms such as ”Cloud Computing” which has not yet been recorded in current existing ontology (e.g., Wordnet).

D. Related Service and Platform

There have been numerous developments in references management tools makes a great assistance in building personal library. Zotero [33] covers thousands of sites that senses content automatically to allow users have a personal library. CiteULike [32] and Connotea [30] provide online service for managing and discovering scholarly references as well as personal library building. Several desktop softwares, including Biblioscape, EndNote, Mendeley and BibDesk with the similar abilities.

Hoang *et al.* [13] proposed a bowser extension and web service called ***Scholarometer*** for academic impact analysis. Also an alternative named *Publish or Perish* [34] that is a software to retrieve and analyze citation information from Google Scholar to present several academic statistics. Their common characteristic is that the individuals of which they utilize the academic citations from Google Scholar to present a fast way to obtain the impact analysis in many aspects. However, users without the rights to make modification to these presented data such as remove the publications which not belonging to self permanently.

Some systems such as Arnetminer [27] and Microsoft Academic Search [26], are committed to provide extensive search and mining services for scholarly communities and network. Both of them make efforts in aggregating academic data from multiple sources and automatically build profiles. In addition, Odysci [28] pays attention to contribute a place where can express opinions on articles. Moreover, Google Scholar Citations [29] provides an easy way for researchers to realize the citation number of the publications.

Based on the analysis of existing state-of-the-art work, it appears that a sophisticated platform with services, such as Key paper searching, Key author localization, is important. In the next chapter, we propose a Web 2.0 platform, *DLC*, which provides several essential services for the scholarly use. There are several distinctions of *DLC*firstly offers a simply way to users to create their own citation repository; DLC recommends key papers and impact scholars by analyzing user's expertise. We also adopt crowdsourcing to take responsibility to maintain the correctness of publications and profiles as well as *Wikipedia*, all the registered users are allowed to edit the publication records on the system.

III. DLC: Digital Library Connector

In this section, we describe the main features of proposed *DLC* platform, which is currently accessible at http://dlc.iis.sinica.edu.tw. The framework of DLC is composed of three conceptual modules: data integration, scholar recommendation and expertise exploration. Figure 1 presents the architecture of DLC. The back-end database is integrated of existing online scholarly repositories which are Google Scholar, DBLP, and user-provided data, such as, researcher's personal publication web page. Digital libraries such as DBLP and CiteSeerX have provided web-services to provide data and metadata, they dump the database into XML format and is compliant with the Open Archives Initiative Protocol. We aggregate attributes (i.e., metadata) from different data sources for different uses. In the *DLC* interface, *DLC* offers various information of author, expertise, publication list, coauthor, related venue, and citation number. To provide the information, several components are designed to fulfill the needs. We elaborate the main components “Data Collection”, “Object Aggregation”, and “Object Recommendation” in the subsections.

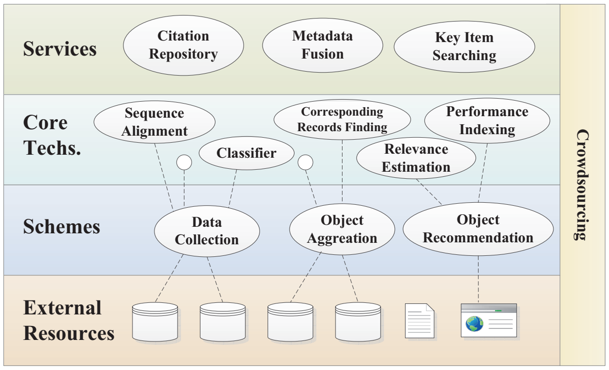


Figure 1. System Architecture

A. Data Collection

In order to improve the satisfactory coverage, DLC integrates Google Scholar, DBLP, and Citeseer. The integrated data are refined and integrated to provide further implementation and services, such as direct answer. We apply the object schema presented by Nie *et al*. [19]. We also employ entity-attribute-value data model to refer the real world entities and relationship among objects based on connectivity. This part is considered the following aspects, including publication data collection, object aggregation, and maintenance scheme. These components and their usage are described below.

This phase commences establishing publications for each scholar by conducting an import task of DBLP XML records, which is formatted and organized bibliographic entries on major computer science venues and available online [23]. *DLC* periodically synchronizes the data from the original data sources, these data includes publication list and citation number. So far we have construct a skeleton structure of our repositories. Each entity extends the metadata (i.e., abstract, reference) from CiteSeerX [25] and Google Scholar [24].

However, in our own observation, the coverage of publications among those scholarly repositories is not perfect. Even though Google Scholar has satisfactory coverage of disciplines and citation information, but name ambiguous problem has not yet been solved. Furthermore, the abbreviation policy of publication in DBLP database causes parsing problem. Previous studies [7], [17], [18] show that the citation records are usually using almost the same sequence of HTML tags and under one parent node. We use our previous work for the extraction of citation records [7]. Furthermore, DLC provides function to import personal publication list which is described by well-organized BibTEX file.

The DLC also applied *Citation Record Extractor* (CRE) [7] for users to provide their personal publication list web page. CRE identifies candidate citation patterns within pages in the DOM tree structure, and then filters out irreverent patterns by using a length-distribution-based classifier. Before returning to client, the BibPro [1] is responsible for citation parsing process. It is a sequence-alignment based citation parser designed to extract components of citations in arbitrary formats.

IV. Object Aggregation

A. Object Aggregation

The aggregation of attributes among those physical scholarly repositories has been performed by using the similarity measure on title field of each citation string. *DLC* regards well-known DBLP dataset as our authorized publication material which contains more than 1.7 million publication records and is easy to be stored with tabular format in a DBMS. *DLC* employs *Edit Distance* to locate the corresponding data which is under a given threshold and the year value must be the same. Hence, *DLC* harvests abstract value from CiteSeerX and citation number from Google Scholar respectively. Note that the aggregation procedure requires time *DLC* establishes multiple query strategies to achieve the satisfactory coverage.

Table I
Query strategies for retrieving the corresponding candidates from Google Scholar

Strategy	Query Composition	Mode
Author	author: <*name*> (with quotes)	Online
Title	allintitle: <*title*> (without quotes)	Queue

B. Query Strategies

DLC regards Google Scholar as the metadata source because the coverage of Google is satisfactory on scholarly lecture. However, several constrains make it difficult to access easily, such as the limited number of queries for a single client and the lack of convenient application programming accessing interface. In practice, *DLC* employs Web proxy technology for cross-domain XMLHttpRequest calls in JavaScript to access the external resources. To guarantee a short response time and save the query number, we have proposed the following query strategies as shown in Table I.

In this scheme, system will not raise the whole update procedure for each client's visit but update those authors who have not yet been updated. According to the observation, we find that the most fluctuations of citation number are centralized in top-k cited papers. Therefore, for a specific author without updating over our defined threshold, the $author$ strategy will be triggered immediately for obtaining the approximately citation number. The number of queries range between 1 and 10 according to the name alphabet order. The similar results with [14] show the authors with unusual names have effective and precise return results. However, those who have the overlap part of names are insufficient for corresponding records and are accompanying with the ambiguous problem. Consequently, totally matched is not guaranteed that merely by using *name* strategy.

The remaining unmatched publications will be delivered to the *title queue* for next query. The *title* strategy just take the whole title without quotes as keyword send to the search engine. Each query will perform the *Edit Distance* measurement to the title of the first returned snippet. In the end, we mark these data to remind the clients to check if any typo exists.

V. Object Recommendation

A. Key Paper Recommendation

For a graduate student, to survey and study papers is important. To realize a details of a topic, it is required to read classic papers and recent presented papers. However, the students may not have experience on knowing which the important papers are? So, one of the component of *Object Recommendation* is *key paper survey* which returns a list of classic and new papers for a specific domain [4].

Searching and identifying the key papers can be regarded as a *recursive* process. Users usually run the process for couple times so that they can have part of important papers.

They usually use some possible keywords, review the returned results, modify the keywords, and use a new query in each iteration. As long as one most relevant paper has been focused. Paper-to-paper can be applied to obtain a key paper by using the perspective on *link* to focus on more key papers. Briefly reviewing a paper can obtain information, including the authors, keywords, and references. We obtain a series of papers of a research topic from the same authors. The information of a paper contains categories and subject description, general terms, keywords, references and context, these information can be processed for the future search. Moreover, cross validating the references from these candidate papers can discover other interesting topic and related technologies. The process usually repeats several time so that user can focus on the research topic.

However, scholarly search engines have developed their ranking mechanisms by processing the user-input. The most common way is to use keyword matching method and to calculate the citation number. Based on the results from search engine, user mostly focuses on the top-*k* highly cited papers as their survey materials. This ranking algorithm brings the challenge of how to skip the well-known papers which we have already learned, to identify the novel papers becomes necessary. Generally, the novel papers are usually lack of enough exposure opportunities [4], which means that they usually receive few citations. Our previous work *Citation Authority Diffusion* (*CAD* in short)is deployed [4], *CAD* is a citation-network-based method to discover potential co-relations to reveal the critical papers. Hence, this unit recommends classic papers and avoids cold start problem.

The whole procedure is triggered by a target research paper tr which was provided by the user. The *information collection* module then analyzes the title, abstract, and keywords of the input paper, and generates the key phrases. The key phrases are extracted by using part-of-speech tagging, linguistic filtering, and *C-value* [15] method. The key phrases are regarded as the input for the scholarly search engine in order to collect the related survey materials sm. In order to retrieve the potential papers and to construct the citation network of sm, the *authority propagator* and *believe propagation* method are applied. The most relevant bibliographies could easily be estimated by using the following Equation (1).

$$rel(s,d) = \frac{InDeg(s \mid d)}{InDeg(d)} \tag{1}$$

where $d \in sm$, current paper of survey material, s is one of the siblings (i.e., references) belongs to d. $InDeg(d)$ represents the amount of citation number of d, and $InDeg(s \mid d)$ means the paper s is cited by d's citer. Moreover, the harmonic mean is calculated and is regarded as the threshold to filter the irrelevant candidates, the detailed equation is described as follows,

$$f(d) = \frac{|sib(d)|}{\sum_{s \in sib(d)} rel(s,d)^{-1}} \tag{2}$$

where $sib(d)$ is the reference set of d, and $|sib(d)|$ is the size of $sib(d)$. Only the $rel(s,d)$ greater than the threshold will be added into the citation network. The main idea is that the more common references they shared, the higher correlation they gained. Citation network is regarded as the input of the authority propagator to identify the key publication list. We leverage the *belief propagation* [16] with our potential function named *authority matrix*. The *belief propagation* focuses on initial weight setting and state transition to diffuse the authority as the belief. The belief propagation is based on the Equation (3) and Equation (4) used to infer the probabilities about maximum likelihood state from each paper in citation network.

$$m_{ij} = \sum_{\sigma'} \Psi(\sigma', \sigma) \prod_{n \in N(i) \backslash j} m_{ni}(\sigma') \tag{3}$$

$$b_i(\sigma) = k \prod_{j \in N(i)} m_{ji}(\sigma) \tag{4}$$

In the above formulas, m_{ij} is the message vector sent by the paper i to j and $N(i)$ is the set of papers citing i, and k is a normalization constant. An authority matrix $\Psi(\sigma', \sigma)$ is exploited on prior state assignment for each citation pair to standardize author duplication rate in citation network from 0 to 1. Generally, a paper holds higher authority if it is cited by another paper who is also having high authority. So, authority propagator regards $\Psi(\sigma', \sigma)$ as a diffusion factor and dynamically updates the authority for each pair. Finally, a converged network with certain authority is expected, i.e., the key paper list in novelty aspect.

B. *Expertise Extraction*

Locating a specific researcher's expertise is always an important and essential task among scholar repositories. This information may facilitate people search which addressed on similar research domains. The expertise extraction procedure composed of two stages, including key term extraction and Wikipedia ontology inference. We incorporate our previous results [2] and [8] to extract researcher's expertise by analyzing their publication list.

VI. CROWDSOURCING

In order to guarantee a satisfactory coverage and to collect more author-publication records, *DLC* uses crowdsourcing mechanism to let registered users to edit. As long as more users join the editing task, the less typos and omissions can be avoided. Registered users are allowed to upload their own publications, we also allow their co-authors and registered users to edit the records. This mechanism is as same as in Wikipedia. Therefore, part of meta-data management in our proposal framework is delegated to the public. Some unique social identifier such as OpenID, Facebook ID are recognized to apply for the ownership of a specific or an interested scholar in system to entitle an applicant the right to maintain the records, including author's profile and publication list. We express more details in publication maintenances in the following.

We propose an additional *verification bit* mechanism in each articles whether they be created by crawlers in advance or those be established recently by volunteers. Those users with authorities for management can perform the *lock* and *unlock* operation to the verification bit. A unlocked publication means that every registered user is allowed to modify. As long as a publication list has been verified, the authorized user can lock this citation to avoid the destruction with evil intension. This manner makes it easier on the management task to have the information with the interested publications, statistics of contributors and the number of the uncertified records.

VII. Conclusion

In this paper, we incorporate our previous work and develop a Web 2.0 platform, namely *Digital Library Connector* (*DLC*). *DLC* provides services to facilitate the tedious research processes. Researchers can easily search key papers, key authors and subscribe key author's academic activity. Moreover, researchers can easily construct their Web 2.0 web pages to present their profile, publication list ,and recent academic activity by using the service on *DLC*. The users show their high satisfactory on using *DLC*.

Acknowledgment

This work is partly supported by the National Science Council of Taiwan, under grant NSC 99-2221-E-011-075-MY3, NSC 100-2631-H-001-012, NSC 100-2811-E-001-003 and NSC 101-2631-H-001-006.

References

[1] C.-C. Chen, K.-H. Yang, C.-L. Chen and J.-M. Ho, "BibPro: A Citation Parser Based on Sequence Alignment," IEEE Transactions on Knowledge and Data Engineering, volume 24, issue 2, pp. 236–250, 2012.

[2] C.-Y. Lu, S.-W. Ho, J.-M. Chung, H.-M. Lee and J.-M. Ho, "Mining Fuzzy Domain Ontology Based on Concept Vector from Wikipedia Category Network," Web Intelligence, 2011.

[3] C.-J. Wu, J.-M. Chung, C.-Y. Lu and J.-M. Ho, "Using Web-Mining Approach for Scholar Measurement and Recommendation," Web Intelligence, 2011.

[4] C.-H. Chen, C.-Y. Lu, H.-M. Lee and J.-M. Ho, "Novelty Paper Recommendation Using Citation Authority Diffusion," TAAI, 2011.

[5] J.-M. Chung, C.-Y. Lu, H.-M. Lee and J.-M. Ho, "Automatic English-Chinese Name Translation in Digital Library Management by Using Web-Mining and Phonetic Similarity ," IEEE IRI, 2011.

[6] K.-H. Yang, T.-L. Kuo, H.-M Lee and J.-M. Ho, "A reviewer recommendation system based on collaborative intelligence," Web Intelligence, pp. 564–567, 2009.

[7] K.-H. Yang, S.-S. Chen, M.-T. Hsieh, H.-M. Lee and J.-M. Ho, "CRE: An automatic citation record extractor for publication list pages," Proc. WMWA, 2008.

[8] K.-H. Yang, C.-Y. Chen, H.-M. Lee and J.-M. Ho, "EFS: Expert finding system based on Wikipedia link pattern analysis," Systems, Man and Cybernetic, pp. 631–635, 2008.

[9] K.-H. Yang, J.-M. Chung and J.-M. Ho, "PLF: A Publication list Web page finder for researchers," Web Intelligence, 2007.

[10] K.-H. Yang, H.-T. Peng, J.-Y. Jiang, H.-M. Lee and J.-M. Ho, "Author name disambiguation for citations using topic and web correlation," Research and Advanced Technology for Digital Libraries, pp. 185–196, 2008.

[11] K.-H. Yang and J.-M. Ho, "Parsing Publication Lists on the Web," Web Intelligence, pp. 444–447, 2010.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang and Z. Su, "ArnetMiner: extraction and mining of academic social networks," ACM SIGKDD, pp. 990–998, 2008.

[13] D.-T. Hoang, J. Kaur and F. Menczer, "Crowdsourcing Scholarly Data," Proc. of Web Science Conference: Extending the Frontiers of Society On-Line (WebSci), 2010.

[14] A. Thor, D. Aumueller and E. Rahm, "Data integration support for mashups," Proceedings of the Sixth International AAAI Workshop on Information Integration on the Web, pp. 104–109, 2007.

[15] K. Frantzi, S. Ananiadou and J. Tsujii, "The c-value/nc-value method of automatic recognition for multi-word terms," Research and Advanced Technology for Digital Libraries, pp. 520–520, 1998.

[16] J. Pearl, "Probabilistic reasoning in intelligent systems: networks of plausible inference," 1988.

[17] B. Liu, R. Grossman and Y. Zhai, "Mining data records in Web pages," Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 601–606, 2003.

[18] C.H.A. Hong, J.P. Gozali and M.Y. Kan, "FireCite: Lightweight real-time reference string extraction from webpages," Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, pp. 71–79, 2009.

[19] Z. Nie, J.R. Wen and W.Y. Ma, "Object-level vertical search," Third Biennial Conference on Innovative Data Systems Research, pp. 235–246, 2007.

[20] A. Thor and E. Rahm, "MOMA - A Mapping-based Object Matching System," CIDR, pp. 247–258, 2007.

[21] Keshav, S., "How to read a paper," ACM SIGCOMM Computer Communication Review, 2007.

[22] Digital Library Connector. http://dlc.iis.sinica.edu.tw.

[23] DBLP XML records. http://dblp.uni-trier.de/xml/.

[24] Google Scholar. http://scholar.google.com.

[25] CiteSeerX. http://citeseerx.ist.psu.edu.

[26] MS Academic Search. http://academic.research.microsoft.com.

[27] Arnetminer. http://www.arnetminer.org.

[28] Odysci. http://www.odysci.com.

[29] Google Scholar Citations. http://scholar.google.com.

[30] Connotea. http://www.connotea.org.

[31] Scholarometer. http://scholarometer.indiana.edu.

[32] CiteULike. http://www.citeulike.org.

[33] Zotero. http://www.zotero.org.

[34] Publish or Perlish. http://www.harzing.com/pop.htm.

Towards Agile Role-based Decision Support for OPC UA Profiles

Dirk van der Linden, Maarten Reekmans
Electromechanics Research Group
Artesis University College of Antwerp
Belgium
{*dirk.vanderlinden, maarten.reekmans*}@*artesis.be*

Wolfgang Kastner
Automation Systems Group
Vienna University of Technology
Austria
k@auto.tuwien.ac.at

Herbert Peremans
Active Perception Lab
University of Antwerp
Belgium
herbert.peremans@ua.ac.be

***Abstract*—Interoperability has become a key factor in modern automation systems. In this context, the OPC (Open Productivity and Connectivity) standards are most prominent. The OPC Unified Architecture (UA) provides platform independent industrial communication using Web-based technologies. It also includes a meta-model guaranteeing interoperability not only at the protocol level but also regarding the semantics of exchanged data. The resulting increased complexity of its specification is an entry barrier for small and medium enterprises. It is challenging to decide which parts of OPC UA a specific company needs to implement. This paper proposes a mechanism for determining the most relevant OPC UA profiles in a specific application domain.**

Keywords-Automation; OPC UA; Profiles; Decision support; Survey; Software tool.

I. Introduction

In the last 10 years, industrial communication has become a key technology in modern industry. A continually growing number of manufacturing companies desire, even require, totally integrated systems. This integration should cover electronic automation devices such as Programmable Logic Controllers (PLCs) and microcontrollers as well as Human Machine Interfaces (HMI) and supervision, trending, and alarm software applications, e.g., Supervisory Control and Data Acquisition (SCADA) and Manufacturing Execution Systems (MES). Industrial communication encompasses the whole range from field management to process management and Enterprise Resource Planning (ERP) applications (business management).

Likewise, the past decade has seen a push towards the integration of building services and building management. Total integration in this field should not only encompass Direct Digital Control (DDC) and SCADA/Building Management Systems (BMS), but also Computer Aided Facility Management (CAFM) applications and HMI ranging from dedicated panels to Web-pads and visitor guidance systems.

The OPC Foundation started in the nineties to promote cross-vendor interoperability for automation projects. Initially, the OPC specifications focused on the Microsoft's proprietary DCOM communication technology. The more recent standard family, OPC Unified Architecture (UA), is designed to be more generic, abstract, technology independent and platform agnostic [1], [2]. However, the resulting increased complexity of OPC UA is an entry barrier for Small and Medium Enterprises (SMEs) that integrate automation systems or provide UA software. The scalability of OPC UA enables interoperability in various fields of application, but it is challenging for SMEs to decide which parts of OPC UA they should implement. The typical activities of SMEs target a specific market niche. In contrast, OPC UA has an almost unlimited field of application. This research project wants to facilitate the process of identifying subsets of the enhanced OPC UA standard for specific targets. Currently, the choice of UA profiles is pragmatically made according to the implementor's knowledge of the OPC UA specifications and the perceived requirements of the application. New concepts of OPC UA, for example information modeling, redundancy or events tend to be skipped due to a lack of awareness of these profiles. Its ultimate goal is a software tool that converts a company's needs to a list of recommended OPC UA profiles, ordered according to the benefits for their business.

This paper gives a short overview on OPC UA and its profiles, then focuses on two major issues concerning OPC UA. First, stakeholders claim that there is a lack of documentation. Second, the generic, abstract concepts of OPC UA result in complex specifications. To discover the most popular environments and technologies and relate the abstract specifications to applications, we performed a worldwide survey. The results of this survey are discussed. In addition, a strategy and tool to improve the choice of profiles are introduced.

The research presented was accomplished within the project "Web-based Communication in Automation" (WebCom), a Consortium Type Project within the EraSME funding program uniting representatives of Research and Technology Organizations and several local SMEs (vendors, system integrators, system developers, consultants). To provide first hand knowledge and expertise on the upcoming OPC UA specifications, the OPC Foundation also agreed to actively support the project and thus serves as an additional partner.

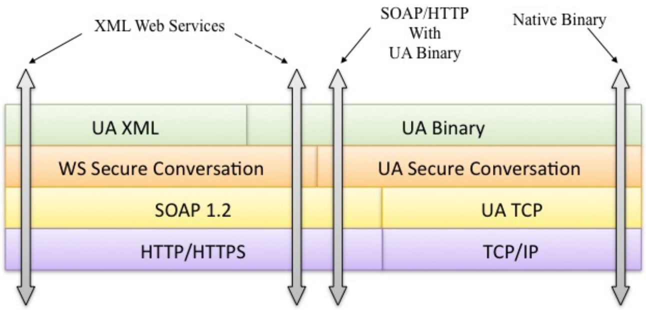


Figure 1. OPC UA transport (Source: OPC Foundation)

II. OPC Unified Architecture and Profiles

Web-based technology is the key to taking interoperability to a new level. Web Services (WS) are totally platform independent – they can be implemented using any programming language and run on any hardware platform or operating system. Easier than ever before, components can flexibly be arranged into applications, collaborating over the Internet as well as corporate Intranets.

OPC UA is considered one of the most promising incarnations of WS technology for automation [3], [4], [5]. From the very beginning, OPC UA was intended as an interface between systems, 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, Distributed Control Systems (DCSs) or DDCs provide another environment for Web-based communication than standard PCs.

The fundamental components of OPC UA are various transport mechanisms and unified data modeling. The transport mechanisms tackle platform independent communication while still allowing 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 volume and firewall friendly transport. As a consequence, two data encoding schemes are defined, named OPC UA Binary and OPC UA XML. Different compromises are possible to find a good balance between security and performance, depending on the application (Figure 1).

Data modeling defines the rules and basic building blocks necessary to expose an information model with OPC UA. Rather than supporting data communication, it facilitates the conversion of data to information. Rather than introducing unnecessary new formalisms, the OPC Foundation encourages definitions of complex data based on related industrial standards. Examples are FDI (Field Device Integration), EDDL (Electronic Device Description Language [6]), IEC 61131-3 (PLC programming languages [7]) and ISA 88 (batch control [8], [9]). Basically, an OPC UA information model is made up of 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 convenient to program against these complex data types. They also bring a potential of code re-use. Note that the OPC Foundation provides dedicated OPC UA information models to structure the legacy OPC specifications (Figure 2). These information models facilitate the migration of legacy OPC interfaces to OPC UA interfaces [10].

OPC UA is designed in a way that individual implementations do not need to support all features, but can be downscaled to a limited scope if desired. At the same time, advanced products which allow a high degree of freedom will require the support of more sophisticated features. A service based OPC UA implementation can be tailored to be just as complex as needed for the underlying application.

Hence, what is needed is a way to describe (and test) which features are supported by an OPC UA compliant product. This is where the OPC UA terms *ConformanceUnit* and *Profile* come into play [11]. A specific set of features (e.g., a set of services or a part of an information model) that can be tested as a single entity is referred to as a ConformanceUnit. An example of a ConformanceUnit is the *Call service*. This service is used to call a method on an OPC UA server. ConformanceUnits are further combined into Profiles. An application (client or server) shall implement all of the ConformanceUnits in a Profile to be compliant with it. Some Profiles may contain optional ConformanceUnits which in turn may exist in more than one Profile (Figure 3). The term *facet* is used to refer to Profiles which are expected to be part of another larger Profile or which concern a specific part of OPC UA. Software certificates contain information about the supported Profiles. OPC UA Clients and Servers can exchange these certificates via services [12].

Up to now, more than 60 OPC UA Profiles have been released [13]. However, it is expected that the list will be extended over time – even by other organizations than the

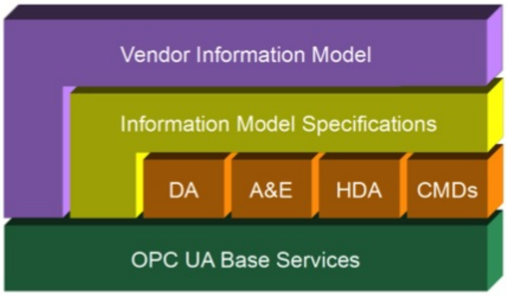


Figure 2. OPC UA Information Modeling (Source: OPC Foundation)

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

OPC Foundation. At this moment, OPC Foundation working groups are working on new, upcoming profiles.

III. Survey

We assume that complaints about OPC UA not being documented well and excessively complex originate mainly from the conflict between the wide range of features and possibilities of the standard on one hand and the strongly focused, niche-oriented business needs of SMEs on the other. Therefore, our goal was to identify which specific sets of profiles add the most value for SMEs. We anticipated that OPC UA implementations for SMEs could be significantly facilitated if the documentation to be provided could be narrowed according to their particular field of application. Consequently, our intent was to determine recommended sets of OPC UA profiles according to the industry sector a business is active in.

This approach follows the assumption that each industry sector requires specific automation applications, resulting in a typical set of automation technologies being used and, likewise, having typical requirements on data communication within and between these technologies. Knowing which OPC UA profile (or combination thereof) is designed to fulfil given communication requirements, it should in this case be possible to recommend a set of profiles based on the industry sector. For example, the redundancy profiles can be recommended for sectors like chemical industry, where high availability is important. Traceability is important for the pharmaceutical sector so the Auditing profiles will be included in the recommended profiles list.

We designed a survey to validate this assumption. The survey did not assume any detailed knowledge of OPC UA profiles on the part of the respondents, but focused on generalized questions regarding communication requirements that would allow drawing conclusions about required profiles. To make sure that these questions reflect the capabilities of the available OPC UA profiles well, we consulted one of the lead authors of the Profiles part of the OPC UA specifications for expert advice. To address a representative number and kind of stakeholders, the survey was distributed to OPC Foundation members as well as companies that figure on the Foundation's regular mailing list and several other industry specific mailing lists containing a wide variety of respondents in addition.

Table I
Industry sectors most relevant to respondents

Oil & Gas Production	18%
Oil & Gas Distribution	15%
Chemical	18%
Food & Beverage	16%
Power Distribution	16%
Power Generation	20%
Building Automation	19%
Automotive Industry	16%
Industrial Automation	38%
Process Automation	30%
IT	19%

Table II
Technologies in use by respondents

ERP	MES	SCADA	PLC	PAC	DCS	TFM	BMS	DDC
30%	24%	66%	72%	29%	43%	9%	18%	13%

About 25,000 questionnaires were sent out, and a total of 719 responses were collected. The geographical distribution of all respondents is shown in Figure 4. It largely matches the geographical distribution of the OPC Foundation members. The most important industrial sectors of respondents (15% or more of responses) are listed in Table I. Multiple answers to this question were allowed. We observed that a significant number (10–15%) of respondents is active in up to 8 different industrial sectors, and still 5–10% are represented in up to 5 areas.

From the results, some conclusions can be drawn as to which parts of the standard are currently of greatest interest and therefore should receive particular attention in general. In particular, we see PLC and SCADA dominating the list of automation technologies in use by the respondents (Table II; PAC = Programmable Automation Controller, TFM = Technical Facility Management). Also, quite general management tasks such as alarm, event and user logging received high importance rankings among respondents.

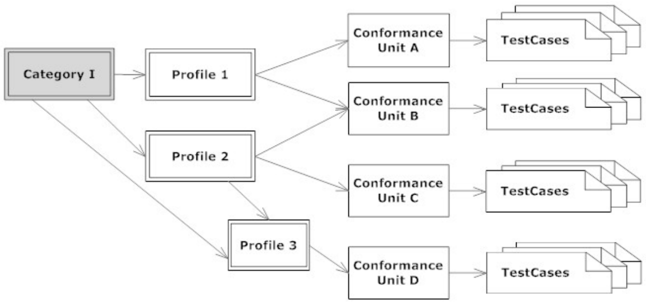


Figure 3. OPC UA Profiles and ConformanceUnits

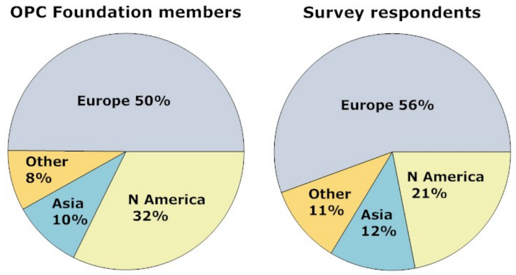


Figure 4. OPC Foundation members and respondents by region

Thus, while there is apparently the need for support of a wide range of different systems, initial implementation effort can be significantly reduced by focusing on these technologies. Considering that many applications follow a very basic pattern, many implementers will only need to provide the so called *Core Server* facet profile, in combination with one *Transport* facet. So an implementation with the same functionality as available with the former DCOM OPC DA, but in a platform independent way. To further reduce the programming effort, wrappers and proxies are available [10].

We found some key trends and assumptions behind the OPC UA technology that can be confirmed by the survey results. For example, 432 interviewees stated to be manufacturer of systems or products that use a communication network. 59% use a field device network, and 37% use the control network in a shared network set-up with the standard computer network. This illustrates the high importance of industrial data communications in general as well as the drive towards combined communication networks and totally integrated systems.

As far as the speed of communication is concerned, communication within less than one second is required by the majority (165/355) of PLC/PAC/DCS users. Also, the time frame for delivering data in the control network is typically short (15% say less than 1 ms; 55% say less than one second). However, a substantial percentage of PLC/PAC/DCS users (81/355) are satisfied with a delivery of data/messages within less than one minute.

This shows that on one hand, demand for fast and efficient transport as provided by UA Binary transport is significant. On the other hand, a large market segment exists where speed does not matter as much as other qualities of service. Also, lower speed may well be acceptable if compensated by other desirable properties such as firewall friendly communication, which would for example be a key property of the SOAP-HTTP WS-SC XML transport facet.

There is also a strong demand for security and robustness. The top three security related issues among respondents are authentication, restricted access and confidentiality of transferred data; for availability, utilizing redundant servers is seen as more relevant than deploying redundant clients.

Regarding operating systems and programming languages in use by the respondents, a technology shift begins to show. Though Windows is still the leading operating system being deployed, a trend towards Linux can be observed. Relevant programming languages are, in decreasing order of importance, C/C++, C#.NET, VB.NET, and Java. The rise of .NET indicates that DCOM is becoming a legacy technology. The use of C#.NET and C/C++ is significantly higher than the other languages ($p < 0.001$). Differences concerning the use of the programming languages in different regions are not significant (at a p-value of 0.05), which leads us to conclude that the technology shift is happening worldwide.

To confirm the suspected dependencies between industry sectors, automation technologies and communication requirements, we applied logistic regression analysis [14] to the survey results. In such an analysis, the estimates of the weight of variables with regards to a specific use provides an idea of the relevance of these variables.

We found the use of MES to be very high in the food and beverage industry. PLC systems are being used nearly everywhere except in power distribution and IT (with negative estimates of -0.54 and -0.89, respectively). The use of DCS systems is also very diverse, except in the automotive industry, which instead shows a significant use of PAC (at an estimate of 0.70). SCADA is present in power generation, industrial automation, food/beverage and oil production, with a negative estimate for the IT sector (-0.64). Overall, PLC and SCADA are quite correlated (0.55).

Again, using logistic regression analysis, we found differences of preferences of programming languages with regard to the type of automation technology in use. The majority of Java users can be found among ERP, MES, SCADA and TFM users (as confirmed by the Hosmer and Lemeshow Goodness-of-Fit test). The majority of C#.NET users work, in decreasing order, with MES, SCADA and ERP systems. The majority of C users focus on SCADA, DDC and BMS systems. The diversity of VB users is the biggest, they work with PLC, SCADA, MES, DCS and ERP systems. The selection of these technologies is based on the analysis of maximum likelihood estimates of a simplified model with an entry cutoff value of 0.15 and a stay cutoff value of 0.15.

Concerning the most common security issues, we found a good fitting logistic regression model showing that ERP users value rogue system detection, auditability of actions, confidentiality of proprietary data and network intrusion avoidance. PLC users have different priorities, with a focus on auditability of actions, availability of systems, restricted external access to proprietary data and network intrusion avoidance. PAC users place a similar (but lower) priority on network intrusion avoidance, availability of systems and restricted external access. MES users assign high importance to preventing the alteration of proprietary data, auditability of actions, network intrusion detection and authentication of users.

The users who need a very short time frame (less than 1 ms) for delivering data/messages via the control network are mostly MES and PLC users. Those who need the fastest message exchange via the computer network (less than one second) are mostly PLC, MES and SCADA users.

We however did not find a straightforward correlation between industry sectors, target technologies and communication requirements that could have been translated into a simple, static set of profile recommendations. We therefore chose to design an agile decision support tool.

IV. Role-based optimization strategy and tool

The logistic regression analysis of the survey results showed that a large number of parameters would be necessary to determine a recommended profile. This makes the regression models too complicated for practical use. In fact, it implies that the correlation between a “typical implementor” (as defined by a naive classification according to target industry sector or automation technology) and any set of recommended profiles is low. It is therefore infeasible to stick to one single set of questions to deal with all stakeholders’ needs. Arriving at this conclusion, we opted for another approach: an online tool to dynamically produce recommended sets of profiles on an individual, user-by-user basis. The tool is based on the generalized questions regarding communication requirements which reflect the capabilities of the various OPC UA profiles that were created for the survey. It is designed to easily accept new or updated questions to reflect newly released profiles. This *agility* is an additional advantage, as the definition of OPC UA profiles by the OPC Foundation working groups is an ongoing process.

Considering our initial goal of identifying which specific sets of profiles would add the most value for a SME, we wanted to have the tool take into consideration the economic dimension in addition to the technical one. Each vendor has its own target market, with a diverse set of customers and specific fields of application. While many profiles might make sense from a technical point of view (and thus may well all be requested by customers), implementing some profiles will provide more commercial benefits than implementing others. Vendors must meet the challenge to find the balance between satisfying customer requirements and return on investment for implementing these profiles.

To best support this decision, our tool should therefore assign a priority to each recommended profile. Also, it should be capable of linking an estimate of commercial benefits based on development time and budget to this prioritized list of recommended profiles. With this information, end-users of the tool can more accurately envision the development planning of a product even without detailed knowledge of OPC UA technologies, as this knowledge is embedded in the tool. Thus, for getting the most relevant results the implementation of the decision support tool takes into account normative constraints (i.e., it shall produce output that is consistent with the OPC UA specifications), budget constraints and maximum commercial benefit (Figure 5).

The decision support tool takes its input from three sources, each representing a particular competence or role. These inputs provide the functional parameters for the decision support tool. When the experts have entered these parameters, the end-user who typically has little knowledge of OPC UA profiles, can use the tool to help determine the list of recommended profiles for their company and application.

The first role is that of an OPC UA expert who is determining the normative constraints. The main task of the UA expert is to input a set of survey questions and possible answers. Each answer is then linked to one or more profiles. Using these relations a profile is produced according to the answer given by a respondent to the respective question. Besides, what we call *static* normative constraints have been hardcoded into the software. Some examples of these static normative constraints are that no product can be built with only one profile and that an application must at least support one of the core facets, one security facet and one transport facet. Another example of a static normative constraint relates to nested profiles: the basic profile must be implemented before an enhanced profile can be implemented (e.g., *Core Server* can only be implemented when *SecurityPolicy - None* has already been implemented).

The second role is that of a software architect who provides input regarding the development time required for implementing a specific profile. The software architect must have detailed knowledge of OPC UA profiles to do this. The development time is put into the tool once per profile. The end-user has to provide some additional parameters like the cost of programming labour in their company, the preferred programming language and an indication of the complexity of the application behind the OPC UA interface to get the total cost of implementation of a specific profile.

Third, the role of technical-commercial manager (sales / business) is to estimate the commercial benefit of implementing a specific profile. This commercial benefit can be estimated and used as a parameter to manage the development priority. Some of the commercial benefits can be estimated by the results of our technology survey. As mentioned, it should be noted that typically the technical-commercial role does not have enough OPC UA knowledge to estimate the benefit of a profile directly, which means that they especially profit from decision support as described in this section.

A survey is restricted to a static, limited set of questions and can never anticipate all upcoming changes and new profiles. In contrast, the decision support software handles every profile, cost and benefit as an abstract parameter. This

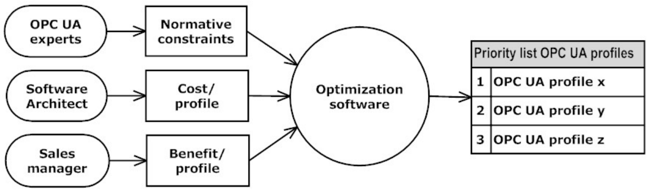


Figure 5. Role-based decision support

allows dealing with diversity and upcoming changes dynamically. New profile cost information and benefit parameters can easily be added.

The online tool is based on CakePHP [15], an open source web development framework. CakePHP follows the Model-View-Controller (MVC) design pattern, which greatly facilitates the creation of database-driven applications such as this. The decision support tool is agile because of the relational database being designed in a way to enable the straightforward addition of:

- New users with a specific role
- Newly released profiles, including their nesting relationship with existing profiles
- New questions and their answers, with an appropriate link to a profile
- New programming languages
- Cost parameters

V. Conclusion and outlook

Thanks to its technology-independence and scalability, OPC UA has a high success potential provided that implementers can reduce the overall complexity by focusing on their specific field of application.

In comparison with the legacy ("classic") OPC specifications, implementing OPC UA is in no way more complex as long as implementers do not attempt to provide more functionality than they need to. On the contrary, it can even be expected that many OPC UA implementations require less effort thanks to the increased development flexibility offered. Formerly, a server implementation had to support all mandatory interfaces, while OPC UA only requires the Core Server facet and a transport facet.

As the conducted survey did not yield a conclusive set of profiles associated with an application field, it was decided that a more personalized result on a user-by-user basis could be more valuable to implementers of OPC UA technology. The agile role-based decision support tool for OPC UA profiles has the goal to deliver a clear cut list of functionality that these implementors need to develop for their products. This will hopefully lead to faster adoption of the OPC UA technology, despite it being generally perceived as involving a very complex set of specifications.

Currently, the framework for the decision support tool is being tested and a working group of OPC UA experts is being assembled to synthesize the questions and answers. By using the input of these experts, a prioritized list of recommended profiles can be produced which is the main feature of the tool and is already implemented. Calculating the cost and commercial benefit is a complex matter and further research will determine the feasibility of these features. When the tool is released after beta testing it will be made freely available to the public in cooperation with the OPC Foundation.

Acknowledgment

This paper is part of the project WebCom [16] funded within the EraSME program by the Flemish Agency for Innovation by Science and Technology (IWT) and the Austrian Research Promotion Agency (FFG). The authors thank all respondents of the survey and the OPC Foundation for constructive collaboration. In particular, we would like to thank Paul Hunkar, one of the lead authors of the OPC UA profiles specification, who has provided valuable consulting expertise. The authors also thank Georg Neugschwandtner.

References

[1] OPC Foundation, OPC UA Part 1 – Overview and Concepts 1.01 Specification, 2009.

[2] International Electrotechnical Commission, OPC Unified Architecture Specification, 62541, 2010.

[3] W. Mahnke, S. Leitner, and M. Damm, OPC Unified Architecture, Springer, ISBN 978-3-540-68898-3, 2009.

[4] J. Lange, F. Iwanitz, and T. J. Burke, OPC – From Data Access to Unified Architecture, VDE-Verlag, ISBN 978-3-8007-3242-5, 2010.

[5] van der Linden D., Mannaert H., Kastner W., Vanderputten V., Peremans H., and Verelst J., "An OPC UA Interface for an Evolvable ISA88 Control Module", IEEE Conference on Emerging Technologies and Factory Automation, 2011.

[6] D. Grossmann, K. Bender, and B. Danzer, "OPC UA based Field Device Integration," Proc. Society of Instrument and Control Engineers Annual Conference 2008 (SICE 2008), pp. 933–938, 2008.

[7] PLCopen and OPC Foundation, OPC UA Information Model for IEC 61131-3 1.00 Companion Specification, 2010.

[8] J. Virta, I. Seilonen, A. Tuomi, and K. Koskinen, "SOA-Based integration for batch process management with OPC UA and ISA-88/95," Proc. 2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA 2010), pp. 1–8, 2010.

[9] van der Linden D., Mannaert H., and de Laet J., "Towards evolvable Control Modules in an industrial production process", ICIW 2011, 6^{th} International Conference on Internet and Web Applications and Services, St. Maarten, pp. 112-117, 2011.

[10] T. Hannelius, M. Salmenpera, and S. Kuikka, "Roadmap to adopting OPC UA," Proc. 6th IEEE International Conference on Industrial Informatics (INDIN 2008), pp. 756–761, 2008.

[11] OPC Foundation, OPC UA Part 7 – Profiles 1.00 Specification, 2009.

[12] van der Linden D., Mannaert H., and De Bruyn P., "Towards the Explicitation of Hidden Dependencies in the Module Interface", ICONS 2012, $7th$ International Conference on Systems, Reunion, pp. 73-78, 2012.

[13] "OPC UA Profiles," www.opcfoundation.org/profilereporting/ (last accessed 2012-01-19).

[14] S. Sharma, Applied Multivariate Techniques, John Wiley & Sons, ISBN 0-471-31064-6, 1996.

[15] "CakePHP: the rapid development PHP framework," www.cakephp.org (last accessed 2012-01-19).

[16] "Webcom Research Project," www.webcom-eu.org (last accessed 2012-01-19).

JASMIN: A Visual Framework for Managing Applications in Service-oriented Grid Systems

Hayat Bendoukha, Abdelkader Benyettou
Department of Computer Science,
University USTO-MB Oran, Algeria
bendoukhyat,aek_benyettou@univ-usto.dz

Yahya Slimani
Department of Computer Science,
University of Tunis El Manar, Tunisia
yahya.slimani@fst.rnu.tn

***Abstract*— Both scientific and industrial applications are becoming more and more complex and need important computing and storage resources to be executed in an accepted time. Workflows associated to service-oriented grids allow to users the specification and the management of their most demanding and interdependent applications. In this paper, we propose a user-friendly framework JASMIN based on a refinement of UML to specify workflow models and on BPEL to generate and compose web and grid services.**

Keywords- Grid computing; Workflow; UML; BPEL; Service composition.

I. INTRODUCTION

Grids are able to aggregate a very big number of distributed and dispersed storage and computing nodes. They support huge databases and execute very demanding applications [1]. However, the most requiring applications in terms of storage and/or computing resources are, in the same time, composed of a set of sub-processes which are interdependent, share the same workspace and have to respect a particular execution scheme to achieve one same objective. Thus, new environments must be able not only to provide all needed resources but also specify, in an efficient way, the internal complexity of users' applications.

As service-oriented technology gains in popularity [2], it is normal that researchers try to design large scale solutions that incorporate web services. Current grids are mainly based on service-oriented architectures developed using grid service infrastructures enabling the invocation of services remotely across Internet [3]. The ability to define, deploy and invoke grid services remotely represents an important barrier for job submission and monitoring, staging, file transfer and data portal services [4]. Indeed, users are involved in many steps of the execution process of their respective applications. Also, in addition to fundamentals and tools of their exercising area, users are constrained to deal with formal languages and complex protocols requiring a very good master of grid technology, web and grid services composition and deployment. This can be noticed by observing the submission process in Globus Toolkit described in the programmer's tutorial of Borja Sotomayor [5].

We consider that it is increasingly necessary to reduce the complexity of the management of service-oriented grids. It is now necessary to associate user-oriented interfaces to large-scale and service-oriented systems in order to hide their complexity and make it easy to handle the services.

The goal of our work is to make grids more efficient and more transparent to individual users by making easy interaction between them and the grid execution environment. In this paper, we propose an approach which links efficiency of service-oriented grids and conviviality of user-friendly composition tools like workflows [6]. We define a workflow and service-based framework JASMIN responsible for submitting and visualizing user applications to a grid system. JASMIN is UML-based for the workflow specifications and BPEL-based for the service composition.

The remainder of the paper is organized as follows. Section II presents some related works and highlights the main contributions of our approach. Section III presents our framework, describes in details its architecture and the functionalities of its components. Section IV concludes the paper and outlines our future work.

II. RELATED WORKS

Workflow has emerged as a useful paradigm to describe, manage and share complex scientific analysis and business processes [7]. Workflows represent, declaratively, the components or codes that need to be executed in a complex application, as well as the data dependencies among those components [6]. Workflow systems address reproducibility by automatically managing the execution of the applications in distributed environments, and by assisting scientists to assemble the workflows and customize them to their particular data. Many researchers are interested, in their projects, to the field of grid computing and workflow [8]. These projects achieved to a variety of management systems for grid workflows, each dedicated to a particular application domain and based on concepts and specific models such as:

- Askalon [9] is a grid application development and computing environment which provides services for composing, scheduling and executing scientific workflows in the grid. Grid workflow applications can be composed using a UML-based workflow composition with Teuta workflow environment [10] or using the XML-based Abstract Grid Workflow Language (AGWL) [11].
- Kepler [12] is one of the most popular workflow systems with advanced features for composing scientific applications. Kepler allows Drag-Drop creation and execution of workflows for distributed applications. Workflows are modelled in MoML (Modeling Markup Language) [13].

ISBN: 978-1-61208-200-4

- Taverna [14] is a collaboration between the European Bioinformatics Institute (EBI), IT Innovation and the Human Genome Mapping Project Resource Centre (HGMP). In Taverna, data models can be represented in XML based language called Simple Conceptual Unified Flow Language (SCUFL) [15].
- Triana [16] is a workflow-based graphical problem solving environment for data mining applications developed at Cardiff University. Triana provides a visual programming interface with functionalities represented by units.

Compared to these related works, our proposal has essentially 3 main characteristics. First, our framework's architecture is not related to any specific application domain contrarily to others like Triana which is dedicated to distributed data mining on grids or Taverna and Kepler which are both oriented to bioinformatics. Second, many frameworks like Triana are based on self defined notations to compose their workflow models. These notations require a learning phase to be mastered and generate models which are difficult to verify and to validate since corresponding toolkits do not provide any verification tool. We avoided these two above disadvantages by using standard tools such as: UML for workflow specification and BPEL for service and workflow composition. We also were widely inspired by the Workflow Management Coalition (WfMC) [6]. Third, our framework is mainly user-oriented. Users interact with our framework through a friendly graphical interface not requiring any specific expertise on formal languages. This can significantly increase the number of grid users. Workflow and UML make our tool much easier to handle than other frameworks that use exclusively XML-based languages as Taverna.

III. THE JASMIN FRAMEWORK

We propose a grid workflow graphical framework composed of two major components. Each component is responsible for one or more specific task in the whole process of management of the distributed application. Our framework JASMIN is the user front-end of the distributed system. It interacts with other service-based and workflow enactment engines in order to accomplish the execution of users' applications. We aim to make grids more efficient and more transparent for different users by making easy interaction between the grid execution environment and the user. Figure 1 describes the architecture of JASMIN and presents its main interactions with the whole system.

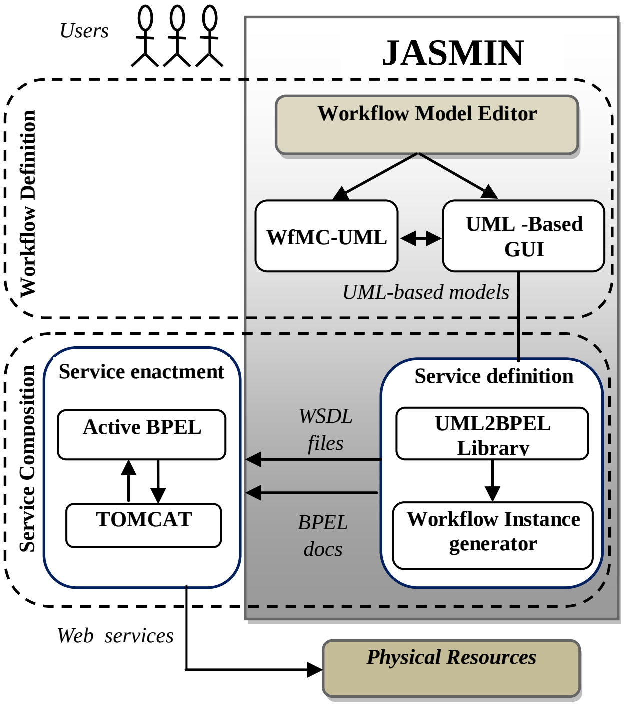


Figure 1. The architecture of JASMIN and its interactions

In order to gain in both efficiency and transparency, we separate between two main steps while composing and deploying processes: the generation of models and the generation of instances. Our framework is workflow-oriented in the first part and service-oriented in the second one. Our architecture is characterized by its capability to separate the specification of applications and their execution. This separation can help to:

- **Easy rewriting of repetitive processes**. Multiple uses of abstract models for a given process are possible without having to redefine it whenever users want to submit the repetitive actions to the grid. Users do not specify conditions on the physical nodes of the grid. This introduces a high degree of transparency.
- **Ease of communication with users**. Users submit their applications in form of diagrams made through a graphical interface easy to handle. This will undoubtedly increase the number of users of grids.
- **Time saving**. Users do not submit sequentially every unit of work separately but realize a global model corresponding to the whole process composed of a set of interdependent activities.
- **Following the evolution of the submitted applications**. Thanks to the graphical interface of the Workflow Model Editor, both submission and visualization of the applications are possible.

In the following, we describe the components of our architecture and their functionalities.

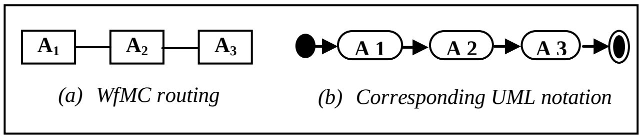
A. *Workflow Definition: The Workflow Model Editor*

The Workflow Model Editor is based on UML activity diagrams. Several editors of UML diagrams exist such as ArgoUML [17]. Our main contribution over these editors is that we focused on both workflow and services. Workflow concepts allow us to see applications through the flow of performed actions. Each application is defined as a set of interdependent activities. The routing rules describe the interdependencies as a control flow of a workflow model and define a formal view of a coordinated set of activities to accomplish the same goal. Besides, our proposal takes into account both the physical constraints of the execution platform and the users' skills. We consider that the execution environment is service-oriented and we suppose that users are not necessary expert and need to be assisted during the submission and the execution of their complex applications.

1) *The WfMC-UML Library*

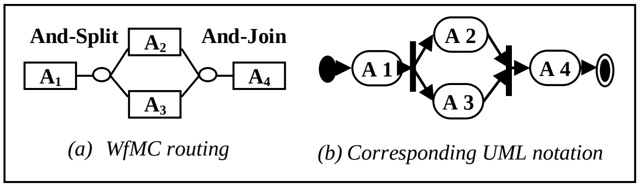
The Workflow Definition tools of JASMIN support the main routing rules defined by the Workflow Management Coalition (WfMC) such as [6]: the parallel routing, the sequential routing, the AND-split, the OR-split, the AND-join, the OR-join and the iterative routing.

Figure 2, Figure 3, Figure 4 and Figure 5 show the most recurrent WFMC routings and their corresponding models in UML activity diagrams.



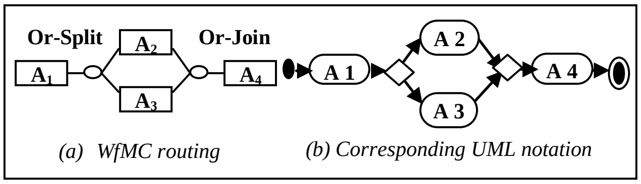
(a) *WfMC routing* (b) *Corresponding UML notation*

Figure 2. The sequential routing



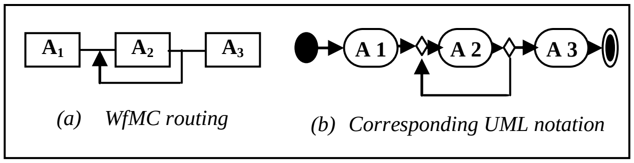
(a) *WfMC routing* (b) *Corresponding UML notation*

Figure 3. The parallel routing



(a) *WfMC routing* (b) *Corresponding UML notation*

Figure 4. The selective routing



(a) *WfMC routing* (b) *Corresponding UML notation*

Figure 5. The iterative routing

2) *The UML-Based GUI*

JASMIN interacts with users through the graphical interface of the Workflow Model Editor. Standard UML formalism does not cover service-oriented applications. We decided to refine UML activity diagrams in order to support two main characteristics of new complex applications which are service-oriented and represent scientific workflow processes [18].

While composing new kind of applications, the importance of workflow concepts (rules, routes and roles) presented by the WfMC changes. In scientific workflows, rules expressing constraints and tasks' characteristics are more important than they are in business workflows. Contrary to rules, roles almost lose sense within new scientific processes since the complexity of applications is no more fixed by the human interactions during the process as in business management but by the interdependencies and the requirements of the activities. In addition, activities of grid workflows are often related to stateful services. Users have to compose a workflow of grid services unlike common workflows of business processes where a workflow of web stateless services is composed [19]. Also, UML-based user interfaces provide, usually, information about activities like name, shared objects, routing rules, dependencies, etc. In order to specify grid workflow applications, our interface provides additional information like activity type, activity communication ports with other activities, etc. Figure 6 shows the main window of the Workflow Model Editor.

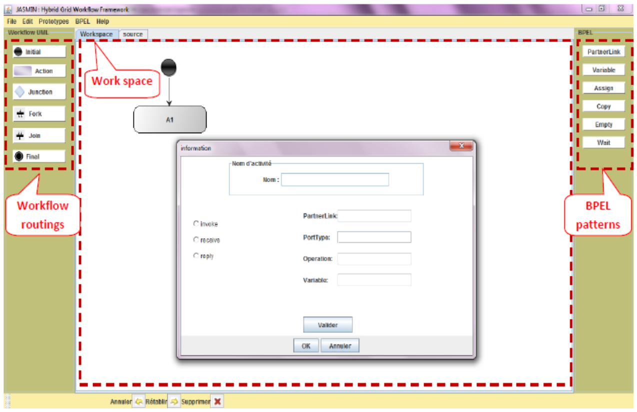


Figure 6. The Workflow Model Editor

Beside the usual patterns, additional toolbars are provided in JASMIN (as shown in figure 6). These ones represent the main refinements that we made on the UML notations. While refining UML, the most important challenge that faced us was: (i) to consider users' skills and provide a convivial interface, and (ii) to consider the service composition language which is BPEL and try to automate the translation of UML models into BPEL documents.

a) Users' skills related refinements: The first type of refinements on UML activity diagrams are related to users' skills. They are defined to ease the work of users and minimize their intervention while submitting and deploying their applications. We consider two different classes of users:

users who are familiar with workflow languages for process management (expert users) and those who have no expertise on UML and workflow (non-expert users).

Expert users represent the class of users from scientific or business fields who already deal with workflow technology and UML formalism. For these users, only standard patterns of UML activity diagrams are requested with no additional patterns or specifications. A tool bar gives access to the main patterns to compose a workflow such as: activity, transition, condition edge, synchronization bar, begin and end nodes, etc. A user can drag any node and drop it to compose a workflow by matching these different patterns together.

Non-expert users are those coming from different application areas and do not necessary have expertise about workflow and UML. However, they need important resources to execute their complex and demanding applications. For this kind of users, we propose a set of predefined prototypes and some dialog boxes to guide them while editing their UML model through JASMIN without forcing them to get a deep knowledge on UML. Users from this class are able to identify the different activities composing the whole process and their interdependencies. Users can use the predefined sub-workflows corresponding to different types of interdependencies to build their whole workflow. Since transparency is being our primary concern, we provide a set of *"prototypes"* corresponding to the most recurrent routings. These routings include, for example, sequential routing, parallel routing (Fork, Join) and selective routing (Switch), as shown in figure 7.

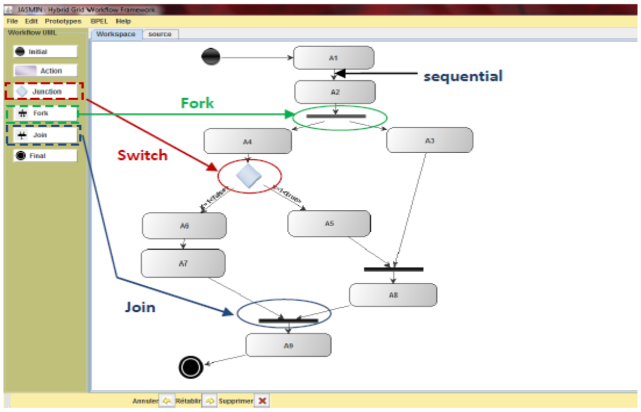
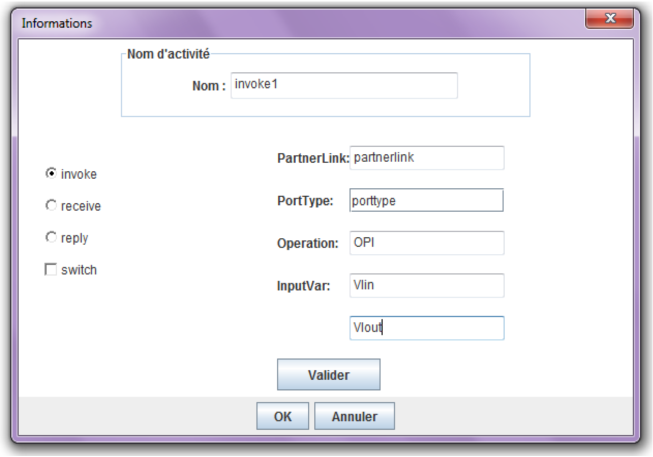


Figure 7. UML prototypes for non-expert users

b) BPEL related refinements: The second type of refinements are related to BPEL notations. Once made, UML activity diagrams have to be managed by service tools. This is possible by enhancing UML notations such as activities, transitions and routing rules with some other patterns like activity properties including, for example, types, variables, port types and partner links. These new patterns are introduced in the UML models in order to make easy the generation of BPEL instances.

Many BPEL patterns are generated automatically, for example, variables and port types in order to reduce users' intervention. However, users still have some informations to indicate like the activity type. While creating any activity, users have to select among a set of activities the type of the activity they wish to insert in the UML activity diagram (see Figure 8). The activity type can be invoke, reply, assign, etc.


Informations

Nom d'activité

Nom : invoke1

● invoke
○ receive
○ reply
□ switch

PartnerLink: partnerlink
PortType: porttype
Operation: OPI
InputVar: VIin
VIout

Figure 8. BPEL activity types

B. *The Service Definition*

The workflow model editor helps to generate the UML models corresponding to a given process. However, even by refining UML diagrams, the execution of the workflow models is impossible, unless we translate these models into services. Service composition tools are responsible for the extraction of the executable jobs of workflow instances in form of services from the initial graphical models. In other words, these tools generate from the UML activities flow a set services written in a formal language. There are many workflow formal languages, but BPEL is the standard for describing the service composition. BPEL contains constraints for control flow and data manipulation as well as interaction activities which model the interaction with web services that implement tasks in a workflow [20].

1) *The UML2BPEL Library*

This library is a set of programs able to generate BPEL tags corresponding to any UML notation in the workflow model. For a portability purpose, the BPEL generation from UML diagrams was divided into two steps.

The first step consists on mapping UML diagrams into Java codes while the second one consists on mapping the obtained Java codes into BPEL documents. This intermediate java code corresponding to the behaviour of the sub-processes and their interdependencies may facilitate a future mapping of UML models into another formal language or creating BPEL documents from other semiformal notation when these ones are coded in java. The second part of the mapping process is from Java code to BPEL document. Each class of the mapping program generates a BPEL activity.

Thanks to the UML2BPEL library and the informations introduced by users while editing UML diagrams, BPEL documents corresponding to both the so-called *basic activities* and the *complex activities* are generated. The basic activities include, for example, invoke activity, receive activity and reply activity. The complex activities represent a set of basic activities grouped by workflow routing rules such as the switch and the flow.

2) *The Workflow Instance Generator*

In BPEL notation, both activities and interdependencies are supported. Comparing to UML, more notions are present in a BPEL definition. As examples, we can mention partner links, port types, variables and activity types [20].

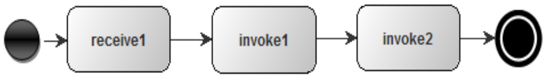
The Workflow Instance Generator is responsible for generating the BPEL documents corresponding to UML models, in coordination with the UML2BPEL library. When users define a new activity or introduce a new pattern related to any activity, the Workflow Instance Generator produces the corresponding code in BPEL. A BPEL document is filled gradually while editing UML diagrams.

Each time a user starts editing a UML activity diagram, a new java file is created. This file is filled while the workflow model is created (when activities are inserted in the diagram or their interdependencies are defined). At the end of the modelling step a Java code corresponding to the whole process is obtained. This mapping from UML to java is invisible to the user.

Beside the generation of BPEL patterns related to basic activities, we also made the generation of BPEL routings automatic. At this level, we proceeded as we did in the generation of the UML models. We implemented some rules to map workflow routing from UML activity diagrams into the so called control flows in BPEL documents. Our grid workflow framework provides a transparent manner to generate the BPEL tags corresponding to the most important WfMC routing rules already presented in the above sections (Sequential routing, parallel routing, selective routing and iterative routing).

3) *An example of a sequential routing in JASMIN*

The workflow definition and the service composition tools of JASMIN allow to generate, respectively, the UML activity diagram and the BPEL document corresponding to a a given application. In Figure 9, we show the UML model of a sequence of three activities: *receive1, invoke1 and invoke2* and a simplified syntax of its corresponding BPEL document as they are produced by JASMIN.



(a) Sequence in UML

```
<process xmlns="http://schemas.xmlsoap.org/ws/2003/03/business-process/"
      name="example"
       …
<sequence>
        <receive partnerLink="PartnerLink" portType="tns:PortType"
                operation="OPR1" variable="VR1" name="receive1">
        </receive>
         <invoke partnerLink="PartnerLink1" portType="tns:PortType1"
                operation="OPI1"  inputVariable="VI1in"
                outputVariable="VI1out" name="invoke1">
         </invoke>
        <invoke partnerLink="PartnerLink2" portType="tns:PortType2"
                operation="OPI2" inputVariable="VI2in"
                 outputVariable="VI2out" name="invoke2">
         </invoke>
 </sequence>
</process>
```

(B) Sequence in BPEL

Figure 9. An example of the sequential routing

C. *The Service enactment*

In order to deploy the workflow as a service in a grid environment, the behaviour description given by BPEL is not enough. It has to be completed by a static description of each activity (service) given in a WSDL file. In fact, the BPEL document shows, for example, which service interacts with which other services and when a given service is invoked. Two kinds of services need to be deployed:

- The web services representing the static description of the workflow and the grid services which are usually deployed on a grid service container, and,
- the workflow services related to BPEL which need a BPEL based workflow engine to be deployed such as ActiveBPEL [21] based on the "Apache Tomcat container".

We believe that it is possible if both kinds of services are deployed on the same service container. We chose to deploy web/grid services on Tomcat instead of the grid container and launch the ActiveBPEL services on Tomcat to allow the deployment of the final workflow services. At this level of our research, we consider that the WSDL documents corresponding to every single involved service available.

IV. CONCLUSION AND FUTURE WORK

Service-oriented grids provide environments to deploy and execute complex applications on distributed and heterogeneous nodes. Despite their performance, Grids stay underweight in terms of ease of use and conviviality. Currently, with the large use of service-oriented technology, workflow tools and languages of service composition are more and more converging. In this paper, we proposed a visual framework for managing applications in service-oriented grids. Our main objective is to take advantage of workflow techniques, service composition tools and grid infrastructures in an easy and transparent way for the user. Our framework JASMIN is based on UML activity diagrams to generate abstract workflow models and on BPEL to generate associated web and grid services to be deployed on a grid environment.

We intend to integrate our framework JASMIN in a service-oriented environment to test physical performances on systems like caGrid [22], Knowledge Grid [23] or Taverna.

ACKNOWLEDGMENT

The authors are pleased to acknowledge Professor Domenico Talia and the members of the GRID Computing Laboratory of the University of Calabria (Italy) for their precious contributions in this research work.

REFERENCES

[1] I. Foster and C. Kesselman, “The grid: blueprint for a new computing infrastructure”, Morgan Kaufman, 2004.

[2] G. Baryannis, O. Danylevych, D. Karastoyanova, K. Kritikos, P. Leitner, F. Rosenberg and B. Wetzstein, “Service composition”, in: M. Papazoglou, K. Pohl, M. Parkin and A. Metzger (Eds.), Proceedings of the service research challenges and solutions for the future internet, Vol. 6500, Springer, Heidelberg, 2010, pp. 55–84.

[3] 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., Technical report, Global grid forum, 2003.

[4] I. Foster, “Globus toolkit version 4: Software for service-oriented systems”, in: Proceedings of the IFIP International conference on network and parallel computing, Vol. 3779, Springer-Verlag, Tokyo, Japan, 2006, pp. 2–13.

[5] B. Sotomayor, “The globus toolkit 4 programmer’s tutorial”, Technical report, University of Chicago, 2005.

[6] R. T. Marshak, “Workflow white paper: An overview of workflow software”, in: Proceedings of the workflow’94 conference, San Jose, 1994.

[7] W. M. P. van der Aalst and K. Hee, “Workflow management: models, methods, and systems”, MIT press, Cambridge, MA, 2002.

[8] J. Yu, R. Buyya, “A taxonomy of workflow management systems for grid computing”, Journal of ACM SIGMOD record, vol. 34 (3), 2005, pp. 44–49.

[9] T. Fahringer, R. Prodan, R. Duan, J. Hofer, F. Nadeem, F. Nerieri, S. Podlipnig, J. Qin, M. Siddiqui, H. L. Truong, A. Villazon and M. Wieczorek, “Askalon: A development and grid computing environment for scientific workflows”, Springer Verlag, 2005, Ch. Workflows for escience, scientific workflows for grids, pp. 450–471.

[10] T. Fahringer, S. Pllana and J. Testori, “Teuta: Tool support for performance modeling of distributed and parallel applications”, in: Proceedings of international conference on computational science, tools for program development and analysis in computational science, Springer-Verlag, Karakov, Poland, 2004, pp. 456–463.

[11] T. Fahringer, J. Qin and S. Hainzer, “Specification of grid workflow applications with agwl: An abstract grid workflow language”, in: Proceedings of the IEEE international symposium on cluster computing and the grid, Vol. 2, Cardiff, UK, 2005, pp. 676–685.

[12] B. Luduscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao and Y. Zhao, “Scientific workflow management and the Kepler system”, Concurrency and computation: practice and experience, Special issue on scientific workflows, vol. 18 (10), 2005, pp. 1039–1065.

[13] A. E. Lee and S. Neuendorffer, “MoML a modeling markup language in XML version 0.4.”, Technical memorandum ERL/UCB M, University of California, Berkeley, 2000.

[14] T. M. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, K. Glover, M. P. Pocock, A. Wipat and P. Li, “Taverna: A tool for the composition and enactment of bioinformatics workflows”, Bioinformatics, vol. 20 (17), 2004, pp. 3045–3054.

[15] G. Hobona, D. Fairbairn, H. Hiden and P. James, “Orchestration of grid-enabled geospatial web services in geoscientific workflows”, IEEE transactions on automation science and engineering, vol. 7 (2), 2010, pp. 407–411.

[16] I. J. Taylor, M. S. Shields, I. Wang and O. F. Rana, “Triana applications within grid computing and peer to peer environments”, Journal of grid computing, vol. 1 (2), 2003, pp. 199–217.

[17] ArgoUML, http://www.argouml.org

[18] M. Sonntag, D. Karastoyanova and F. Leymann, “The missing features of workflow systems for scientific computations”, in: G. Engels, M. Luckey, A. P. and R. Reusner (Eds.), Proceedings of workshops on software engineering, Vol. 160 of LNI, GI, Hanoi, Vietnam, 2010, pp. 209–216.

[19] W. Dou, J. L. Zhao and S. Fan, “A collaborative scheduling approach for service-driven scientific workflow execution”, Journal of computer and system sciences, vol. 76 (6), 2010, pp. 416–427.

[20] D. Jordan, J. Evdemon and A. Alves, “Web service business process execution language version 2.0.”, Technical report, OASIS standard, 2007.

[21] The ActiveBPEL Project, http://www.activebprl.org

[22] J. H. Saltz, S. Oster, S. Hastings, S. Langella, T. M. Kurc¸, W. Sanchez, M. Kher, A. Manisundaram, K. Shanbhag and P. A. Covitz, “cagrid: design and implementation of the core architecture of the cancer biomedical informatics grid”, Bioinformatics, vol. 22 (15), 2006, pp. 1910–1916.

[23] E. Cesario, M. Lackovic, D. Talia and P. Trunfio, “A visual environment for designing and running data mining workflows in the knowledge grid”, In: Data mining: foundations and intelligent paradigms, D. Holmes, L. Jain (Editors), Springer, Intelligent systems reference library, vol. 24, 2012, pp. 57--75.

Local Context Anchoring in the Modern Web Service Retrieval Model

Konstanty Haniewicz
Poznan University of Economics
Poznan, Poland
konstanty.haniewicz@ue.poznan.pl

***Abstract*—Local Context Anchoring is one of mechanisms supporting the modern Web service retrieval model. Its aim is to provide users with support in the retrieval of data on Web service operations that are beyond boundaries of their Suborganisational Units. Due to the fact that user operates outside his regular environment a mechanism is necessary to make up for the lack of certainty on the structure and content of queried resources. In order to present the mechanism in a satisfactory manner a description of the key concepts that are motivation for modern Web service retrieval is given. Their scope and focus is different from the one presented in majority of publications concerning the Web service domain. The model itself is also introduced along with details on its structure and features important to targeted users.**

***Keywords*-Local Context Anchoring; Web service description; functionality description; knowledge representation; Information Retrieval**

I. Introduction

In order to satisfy a constant need for up to date data on available resources a novel model for Web service description is proposed. It takes into account a number of initiatives addressing Web service description and retrieval based both on solutions relying on semantic enhancements ([1]–[5]) and those that mainly employ standard Information Retrieval based techniques ([6]–[10]). The critical analysis coupled with feedback gathered from the Information Technology professionals led to the definition of five key aspects of the desired solution which later became a ground work for the definition of the proposed model.

The description of the key aspects is given below.

- **Effectiveness** - is perceived as the ability to cater for a need expressed by a user in a format provided by the available solution. When IR-based solution is taken into account, standard measures of precision and recall should provide the answer to how effective a given solution is. On the other hand, in the case of semantic based solutions, precision and recall are deceiving because when an ontology is being queried it should always provide a complete set of answers. Thus, a single measure cannot be applied and a description of effectiveness in terms of solution specification must be available.
- **Cost** - is an effort that should be spent on a proper description of a desired artifact with an envisioned technology, time spent on learning necessary description techniques and a prognosis on a timewise performance of analyzed solution. This set of properties penalizes solutions that require a lot of effort in preparation for production and use and, in addition, a long time of query matching. Such bias leads to the promotion of solutions with a low level of complication as perceived by the end users.
- **Scalability** - describes how soon and to what degree a performance shall drop along with an increase in a number of handled Web service operations. This measure shall promote solutions that can handle thousands of Web services with tens of thousands operations.
- **Scope** - involves any important additions to a baseline of a Web service description by Web Service Definition Language (WSDL) documents in terms of an identification of a vital, not previously addressed areas of importance to a user, any extension of description besides functional description of Web service operations, such as business key performance indicators [11] and a perspective on Web services different than that of a developer.
- **Purpose statement** - determines whether an initiative allows stating the purpose of a Web service along with its operations. It is understood as a possibility to express a goal of an artifact in question so that it is clear what it does for all the parties involved.

In order to address all of the enlisted aspects, one could not longer work with the traditionally developed models as proved incapable of delivering satisfying results in the majority of the enlisted aspects. A broader discussion is given in the Related Works section.

Thus, the presented model is not a monolithic in the sense of common ontology that had to be designed, produced and deployed by a relatively small and extremely well coordinated team in order to be successful. What is more, it does away with the idea of indexing and extending WSDL documents in order to treat them as a set of regular textual data with specific metadata.

The postulated model aims for a federation of interlocking term networks ordering terms used in the description of Web services and their operations. These terms are gathered on a small scale, in order to make it possible for a relatively small Suborganization Unit (hereafter refereed to as SU) to exhaustively describe their Web service assets with the vocabulary of their choice that is perfectly understood in SU's context.

The document is organised as follows: first, an overview of the model is given with a special highlight of the mechanism of phrased-based description and its usability for different groups of users; next, a more detailed overview of mechanisms is discussed; following that, a presentation of Local Context Anchoring is given along with validation subsection presenting the experiments' results; summary section is preceded by the Related Works section.

II. Modern Web service description model

The first premise of the presented model is to make a Web service an asset available to a wider range of users in an organization. Thus, its usability is not only based on the mastery of details desired by technology-oriented personnel but also on various features deemed important by business and executive users. The stakeholder environment is depicted in Figure 1. While the three given groups share some needs one has to explicitly emphasize that business users are more interested in Non Functional Parameters (NFPs) and an actual usage of any given Web service operation in projects of interest to them. They should also be more interested in the purpose of an entity that is crucial for their business processes. This is contrasted with the emphasis given by Information Technology professionals on technical details, such as Inputs, Outputs, Preconditions and Effects, Quality of Services and actual service-providing systems. Executive users might be interested in general costs of invocation, compliance with Service Level Agreements (SLAs) and overall involvement of any given Web service operation across an organisation.

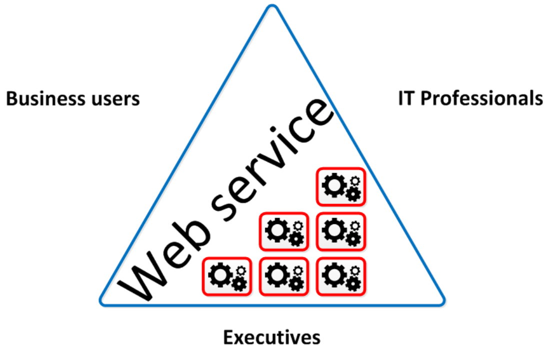


Figure 1. Different interest groups which can benefit from a different approach to a Web service description

The model itself is build around the notion of phrase-based Web service operation description aided by the above-mentioned NFPs to cater for business-oriented queries. The general phrase-based description is structured as follows:

Web service operation: $\langle(\alpha, \beta, \gamma), \mathbf{nfp}\rangle$

- α – action
- β – object
- γ – action-object supplement
- **nfp** – vector of NFP and its values

The decision to mold it in this particular way is derived from the prevalent lack of purpose statement in the majority of the surveyed initiatives. The purpose statement should be understood as a method of answering the question of what a given Web service operation does in a given context. One can argue that a Web service operation name shall convey this information, or even that a definition of Input and Output values in an ontology used throughout the organisation is sufficient.

Unfortunately, one cannot agree with this due to the fact that names ar often poorly defined and that the connection between the purpose of a Web service operation and its input and output parameters is somewhat remote. Even if when the Web service purpose is stated as a goal encrypted as a series of references to an ontology one can easily check that such definition is unfathomable to an average business user [12].

Instead, the phrase-based description builds on the federated effort of SUs that should possess sufficient knowledge to catalogue their Web service assets with terms and phrases deemed most suitable in their context. The process of cataloguing is semi-automated as the model is able to foresee tools that should accept a number of documents which are to be treated as a reference material to obtain an initial list of important terms that might be included at a later phase.

When accomplished, Local Controlled Vocabulary (LCV) serves as a master list of terms and phrases in a given SU. Web service operations are described with terms originating from LCV. It is very important as there is no guarantee that, when any given SU is preparing its LCV, a number of terms or compound terms used in descriptions will not be repeated. In addition, namespaces allow for customizing the results of Web service retrieval and mapping of terms across an organization.

A syntax of the phrase-query language is given below in the form of Antlr [13] grammar, where actual terms used to denominate phrases and namespaces were substituted by exemplary ones in order to make the syntax brief:

```
grammar phrase_query_grammar;
phrase_query: (a b g) nfp*;
a : 'a:' namespace compound_term;
b : 'b:' namespace compound_term;
g : ('g:' namespace compound_term)+;
namespace : '#' ('aaa'|'bbb'|'ccc');
compound_term :  term+;
term   : 'aaa' | 'bbb' | 'ccc' | 'ddd';
nfp : nfp_el ':' val;
nfp_el : 'nfp1' | 'nfp2' | 'nfp3';
val : sign number;
sign   : '+' | '-' |;
number  : digits '.' digits;
```

```
digits  : ('0'..'9')+;
```

III. Local Context Anchoring

Thanks to the design decisions reported in the previous section, a user faces three possible ways of interaction with the repository built upon the postulated model:

- query based on phrases,
- Web service operation name lookup,
- free query.

The most complicated case is the free query as system has no control over what a user inputs thereinto. What is more, due to the federated nature of the model this is the key scenario when various LCV become integrated.

When querying for a term, a list of matching resources, ordered by one of the phrases, is given. When there is no match, a set of hints is presented based on term references with the employed knowledge representation structure.

The mechanism for Local Context Anchoring (LCA) depends on open data repositories or assorted organisation corpora. In its essence, LCA can be viewed as a specialised Query Expansion algorithm [14] that takes into account organisation specific data.

There is a wide choice of data resources that can be adopted by the individual SUs and organisations as a whole [15].

An organization can possess vast resources on some topics that are key to its objectives and cannot be matched by those available in the open repositories.

The key concept of Local Context Anchoring is probing the available repositories (of open access type or propriety) for terms that coincide with those unmatched with terms used in the descriptions of the Web service operations. These are not to be understood as traditional measures used by Information Retrieval based on statistics of direct neighbour co-occurrence. The unmatched term is queried across the available resources. When matched, its context is probed for terms present in a set of all defined terms across all possible namespaces in addition to a check of the actual descriptions of Web service operations. The context of a search term is understood as a frame that spans for n terms before and after the matched term. The actual number of terms is dependent upon experiments, yet the research performed demonstrates that the frame which matches the length of an average paragraph in English texts is a good choice (100 to 150 terms).

The terms that fall into the frame are normalized and stop words are removed. The mechanism yields best results when multiple matches are found and an the occurrence ranking of coinciding terms can be prepared at a later stage of this algorithm. The retrieval of coinciding terms makes sense only in a situation, where resources responsible for providing the context are rich enough.

A. Auxiliary elements of LCA

The above described mechanism is made more adaptable by the inclusion of local resources that are invulnerable to network latency problems or other access issues.

Best example of the desired kind is Wordnet [16]. Moreover, the ability to include terms more specialised, less specialised and a variety of synonyms enriches the set of possible hints. More, the terms resulting from coinciding term search obtain a higher score if they are matched by the terms obtained from resources such as Wordnet and the like (The Suggested Upper Merged Ontology - SUMO [17] which is integrated with Wordnet or ResearchCYC [18]).

At the time of preparation, a prototype uses the most important open repositories, Web search engine, Wordnet and SUMO version integrated with Wordnet along with a number of resources compiled in such a way that specific domains are better represented.

All of these auxiliary measures are introduced to make the model respond to a user's needs to the furthest possible extent. The model is built upon a presumption that failure to present an answer is the worst case scenario which should be avoided at all costs.

Thus, the mapping across functionalities described in various SUs is used. As discussed above, a SU has a perfect freedom of choice when it comes to a set of terms related to its needs and its particular business environment. Yet, many a time, a situation can occur in which some aspect of functionality was described with a term that had many used synonyms in other LCVs. LCA uses this as an opportunity to draw a set of mappings across various SUs in order to come up with yet another data source that could be used when a free query is to be answered.

As there is no guarantee that entities described with similar terms have similar functionality due to the above emphasized reasons, a decision on similarity has to be made by a query issuing user.

LCA and its auxiliary submechanisms can generate a lot of new possibilities, therefore some restrictions had to be introduced. Throughout the experiments with randomly generated descriptions it became apparent that the search for coinciding or similar terms should be restricted to the first two phrases (α and β). One has to remember that each possibility for α is checked with every possibility for β phrase, which results in quadratic complexity of the whole pass.

The decision was made to avoid exponential expansion of the problem space, especially that γ phrase had no limit on the number of the used terms (although a close inspection of the available Web services can provide grounds for a reasonable arbitrary limit). An attempt to match a term that is unknown a priori and arrives for processing as an outcome of the Local Context Anchoring mechanism resembles efforts of the automatic matching of Semantic Web services. Service composition was proved to be feasible, yet the performance of this procedure drops significantly with the increase in the number of Web services [19]. Hence, the above constraints.

The whole mechanism can be outlined in the following steps:

- A query is issued by a user;
- The operation is handled by Local Context Anchoring:
  - each term is submitted as a query to open resources;
  - each term is submitted as a query to the available term networks;
  - each item is consulted with the available mappings for Local Controlled Vocabularies;
- The LCA computes preliminary lists of hits from queried resources;
- A List is ranked with the frequencies of hits in given resources;
- A List is modified to satisfy the predefined trust levels associated with resources;
- Matching of ranked terms with the available Web service operation descriptions is performed;
- A List is rectified by the reordering of Web services depending on a level of match (where compound terms are scrutinized);
- A final output is presented to a user.

As a feature of user interface, results can be displayed as a flat list or grouped by a variety of criteria, such as a project to which the matched Web service operations pertain, a namespace of home LCV or a Web service with which it was deployed.

B. Compound term decomposition

In order to balance flexibility and expressivity of the solution, the transformation of singular terms into compound ones was allowed. This is clearly visible in the previously presented syntax of a phrase query.

Compound terms were introduced to prevent the unwanted decrease of the solution's performance as terms built with other terms could be easily implemented. It is only an addition of extra layer of abstraction that allows for storing base terms that does not negatively affect the benefits of the Local Context Anchoring.

Hence, all the compound terms must be built with a tool that stores data on atomic terms used to produce a compound.

Having accomplished this, a reconstruction of terms is a simple procedure that looks a compound term up in a designated register. Users benefit from this feature as they can easily forge new description phrases with terms that suit the character of their SU best. On the other hand, the solution does not lose effectiveness in situations, where a direct match is not present in the repository. Thanks to the discussed features of LCA the Web service operations described even with the most complicated compound terms can still be matched.

C. Result caching

Query caching in Information Retrieval has a very important role as a feature that boosts effectiveness of any IR system [20].

The proposed model also includes mechanisms that allow for the results of queries to be cached. This is mainly dictated by the need of further efficiency gains in terms of execution time. The novelty of the caching mechanisms proposed here is based on the retention of cache data throughout the system's lifecycle. It can be achieved mainly due to a relatively small amount of data concerning Web services. This is to be contrasted with astronomic amounts of data concerning documents in the process of being indexed by common purpose Web search engines.

The essence of this mechanism lies in the fact that every description of a Web service operation is retained in the solution implementing the model and the terms used in the phrases are easily traceable to the previously issued queries. Therefore, once a query has been issued and a set of operations has been retrieved, it should be constantly updated so that the subsequent queries referring to the initial result set could bypass the initial mechanism. This short-circuits the whole process, cutting drastically the number of operations required in order to present the answer to a user.

In order to prevent memory exhaustion, a set of supporting mechanisms is implemented. Such auxiliary mechanisms allow for the coordination of the caching mechanism with the frequency of a particular query. When some previously defined threshold is met, a query result set is cached.

The introduced modification of caching mechanisms satisfies not only the effectiveness, but also a the cost and the scalability. Given that data in cache always cover the complete set of the matching Web service operations, one does not risk a lapse originating from the fact that some important operation was omitted.

The scalability aspect is reinforced by decreasing the size of the search space by the introduction of mapping between frequently issued queries and their constantly updated result sets. What is more, there is no penalty from using cache as there is no possibility of having it outdated.

D. Validation

In order to ensure that the postulated solutions yield desired results a number of experiments was performed. A quality that might be crucial to many, is the validation of effectiveness measured in terms of execution speed of the algorithm responsible for matching user-issued queries against the available descriptions. The experiments were conducted in a moderately fast test bed consisting of a workstation equipped in Pentium Core 2 Duo (Allendale - 2.4 Ghz) processor with 3 GB of RAM. The algorithm was implemented in the Python programming language (tests run with PyPy implementation version 1.7).

The test scenario was centered on a repetitive retrieval of Web service operation descriptions matching the prepared query. The initial analysis of the solution proposed in the model indicates that the matching process has the worst complexity of $O(n * m)$, where n and m are the numbers of elements from the matched sets.

The Web service operation descriptions used were generated in an automated manner. All of the tests started with as few Web service operation descriptions as 100 and gradually increased up to one million and finally to 10 millions. The upper bound was dictated by the amount of the available memory in the test environment. Every matching operation was performed 100 times with different queries. The initial experiments with the first version of the matching algorithm proved that it was feasible to apply it to the general task. The summary thereof is given in Table I.

Table I
RESULTS FROM EFFECTIVENESS EXPERIMENT MEASURING EXECUTION TIME OF QUERY MATCHING ALGORITHM

Number of descriptions	Average time after ten runs (seconds)
100	0.000053912401
1000	0.000157743692
10000	0.001043230295
100000	0.007972598076
1000000	0.103747159243

Table II
RESULTS FROM EFFECTIVENESS EXPERIMENT MEASURING EXECUTION TIME OF IMPROVED QUERY MATCHING ALGORITHM AGAINST TEST DATA REFLECTING THE NEW STRUCTURE OF WEB SERVICE OPERATION DESCRIPTION

Number of descriptions	Average time after ten runs (seconds)
100	0.000038175809
1000	0.000092013316
10000	0.000257968902
100000	0.001834869384
1000000	0.021018028259
10000000	0.441106071472

The test programme was implemented as a single threaded application. The original run included the following test scenario data:

- 90 α phrase elements to chose from,
- 90 β phrase elements to chose from,
- 90 γ phrase elements to chose from,
- 190 nfp elements to chose from.

As an additional constraint a query had at most 1 alpha phrase element, 1 beta phrase element, from 3 to 7 gamma phrase elements and up to 5 nfp elements. As one can seen, the initial version performed well until a certain number of Web service operation descriptions was to be handled. Since results deviated from the initial assumptions, the code was scrutinized, optimized and extended in order to accommodate additional features.

What is more a hypothesis that Web service operation names should not convey so many phrases was formulated. In order to validate it a survey was conducted with the help of data gathered in the course of the research activities (a corpus of over 50000 WSDL documents). The WSDL documents selected as a means of hypothesis verification were characterized by the highest number of operations. Of course, service architects responsible for the analyzed Web services did not use the function denomination method presented in this work. Fortunately, a number of scrutinized Web services had an easy-to-follow scheme of names, where one could easily distinguish between functional objectives of the terms used for the description. The outcome of the verification can be summarised by the following structure (it is approximated from the eligible WSDLs).

- 50 α phrase elements to choose from,
- 90 β phrase elements to choose from,
- 110 γ phrase elements to choose from.

What is more, one had to arbitrarily establish a number of validated NFPs that should be taken into account. The rationale for this is the fact that every organization is interested in providing of the minimal common set of Key Performance Indicators that makes it possible to compare results globally. The minimal set of 5 enlisted NFPs was used, namely Cost, Availability, Reliability, Performance and Security [21].

It is important to notice, that it is difficult to clearly distinguish between β and γ phrase elements outside the model presented here, where it is tracked throughout the whole process, even if it happens that one term can be used in both phrases, while being a member of different namespaces. Therefore, the above given structure conveys some possible redundancy of terms used in the discussed sets of phrases – their intersection is not empty.

Table II summarizes the results. One can see that the effectiveness was greatly improved in comparison to the initial experiment. The presented implementation handles queries on the body of 10 million descriptions easily and it was measured that it could maintain its effectiveness for a larger corpus by parallelization of the whole procedure. Thanks to the low cost of merging individual sets (an operation of adding a set to another set has a complexity of $O(n + m)$, where n and m are the numbers of elements from summated sets) and cheap partition of corpora into smaller chunks the overhead on the parallelization is very low.

The obtained result data allow one to conclude that the proposed method is effective in terms of execution speed and it scales well up to 100 million descriptions (there were no additional experiments above that number). As mentioned previously, the additional features were introduced in the improved version. While matching the query against the Web service operation description not only full matches are given, but also those that use the desired terms originating from other namespaces.

IV. Related work

The model and mechanisms presented here are an answer to various issues and challenges raised in the literature on the Web service description [22]–[25]. Historically, the two most important approaches were the Information Retrieval and semantic annotations. If one is to compare the two, it can observed that they are not mutually exclusive as to their global goals but they differ significantly when it comes to the means to present a user with the desired functionality.

In general, the IR-based methods treat WSDLs as text documents, where the desired terms either exist or not. The actual implementation can differ significantly from the common border line by the deployment of various techniques that try to avoid simple term presence verification by virtue of thesauri or other similar means. Nevertheless, the success of a user query is strictly connected to one's choice of query terms. The most important works representing this approach are [6]–[10] .

On the other hand, the semantic-based solutions focus on functionality description with advanced description languages that can perfectly describe user needs in terms of common ontology of concepts. This results in a perfect match of descriptions against user queries with a number of important side effects. This approach is best represented in [1]–[5]. Of importance is the fact that the most precise semantic mechanisms are prone to deficiencies of the cost nature [26].

Apart from the two most important approaches mentioned above, there are solutions that try to leverage their best traits. The means by which the hybrid solutions try to achieve their goals are very different for each case and cannot be summarised successfully for the whole population. The most influential initiatives include [25], [27], [28].

It was decided that each group of the solutions should be ranked against the five key aspects mentioned in the introductory section of this work. The result is available in Table III. The referenced works are only a fraction of those used for the comparison presented below. The total body used was composed of 44 positions (from over twice the number initially considered). All the aspect except Cost should be read as the higher percentage the better, whereas Cost should be understood conversely. The percentage is calculated on the basis of the individual evaluation of the works classified according to the above-mentioned approaches. There were 12 works classified as a pure IR-based approach, 22 classified as the semantic-based approach and 12 were classified as the hybrid approach.

Table III
Quantification of key aspect conformance according to domain literature

	IR based	**Semantic**	**Hybrid**
effectiveness	$\approx 50\%$	$> 75\%$	$\approx 50\%$
cost	$> 50\%$	$< 25\%$	$\approx 50\%$
scalability	$\approx 50\%$	$> 25\%$	$\approx 25\%$
scope	$< 25\%$	$< 25\%$	$\approx 50\%$
purpose	$< 25\%$	$\approx 25\%$	$\approx 50\%$

V. Conclussions

The work presented introduces a flexible and effective Web service description model that, thanks to its structure and features, minimizes the cost of Web service description and allows for a better scalability of the solution. The cost is decreased by the simplification of the description process aided with compartmentalisation of an organisation into sufficiently small units with a clearly defined context used with reference to its Web service assets. The scalability is achieved by resigning from the fully-fledged semantic description that relies on reasoning subsystems characterised by superb precision, yet suffering from low timewise efficiency.

The most important mechanisms were implemented as a prototype in order to measure the effectiveness of various mechanisms. The gathered data demonstrated that query matching with phrased queries was feasible for tens of millions of Web service operations in much less than one second. It was achieved thanks to the structure of query and ability to effectively filter off the unnecessary Web service operations. Local Context Anchoring was initially tested and proved to fetch results in more than 90% of test cases. Yet, the results must be tested manually by users in order to measure their satisfaction with the provided answers.

References

[1] P. Hitzler, M. Krtzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "Owl 2 web ontology language primer," *W3C Recommendation*, vol. 27, no. October, pp. 1–123, 2009. [Online]. Available: http://www.w3.org/TR/2009/REC-owl2-primer-20091027/

[2] D. Roman, U. Keller, H. Lausen, J. D. Bruijn, R. Lara, M. Stollberg, A. Polleres, C. Feier, C. Bussler, and D. Fensel, "Web service modeling ontology," *Applied Ontology*, vol. 1, no. 1, pp. 77–106, 2005. [Online]. Available: http://portal.acm.org/citation.cfm?id=1412350.1412357

[3] I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosof, and M. Dean, "Swrl: A semantic web rule language combining owl and ruleml," *Syntax*, vol. 21, no. May, pp. 1–22, 2004. [Online]. Available: http://www.w3.org/Submission/SWRL/

[4] R. Akkiraju, J. Farrell, J. Miller, M. Nagarajan, M.-T. Schmidt, A. Sheth, and K. Verma, "Web service semantics - wsdl-s," *International Business*, vol. 2008, no. Version 1.0, pp. 1–42, 2005. [Online]. Available: http://www.w3.org/Submission/WSDL-S/

[5] J. Kopecky, T. Vitvar, C. Bournez, and J. Farrell, "Sawsdl: Semantic annotations for wsdl and xml schema," *IEEE Internet Computing*, vol. 11, no. 6, pp. 60–67, 2007. [Online]. Available: http://oro.open.ac.uk/28624/

[6] J. Zhou, T. Zhang, H. Meng, L. Xiao, G. Chen, and D. Li, "Web service discovery based on keyword clustering and ontology," *2008 IEEE International Conference on Granular Computing*, vol. 1, pp. 844–848, 2008.

[7] M. Bruno, G. Canfora, M. D. Penta, and R. Scognamiglio, "An approach to support web service classification and annotation," *2005 IEEE International Conference on eTechnology eCommerce and eService*, pp. 138–143, 2005.

[8] B. Srivastava, K. Ponnalagu, N. C. Narendra, and K. Kannan, "Enhancing asset search and retrieval in a services repository using consumption contexts," *IEEE International Conference on Services Computing SCC 2007*, no. Scc, pp. 316–323, 2007.

[9] M. Espinoza and E. Mena, "Discovering web services using semantic keywords," *2007 5th IEEE International Conference on Industrial Informatics*, vol. 2, pp. 725–730, 2007.

[10] H. Gao, W. Stucky, and L. Liu, "Web services classification based on intelligent clustering techniques," *2009 International Forum on Information Technology and Applications*, pp. 242–245, 2009.

[11] B. Marr, G. Schiuma, and A. Neely, "Intellectual capital defining key performance indicators for organizational knowledge assets," *Business Process Management Journal*, vol. 10, no. 5, pp. 551–569, 2004. [Online]. Available: http://www.emeraldinsight.com/10.1108/14637150410559225

[12] M. Klusch, *Semantic Web Service Description*. Birkhuser Basel, 2008, vol. 16, no. 2, pp. 31–57. [Online]. Available: http://www.springerlink.com/index/10.1007/978-3-7643-8575-0

[13] T. Parr and K. Fisher, "Ll(*): the foundation of the antlr parser generator." in *PLDI*, M. W. Hall and D. A. Padua, Eds. ACM, 2011, pp. 425–436.

[14] E. M. Voorhees, *Query expansion using lexical-semantic relations*. Springer-Verlag New York, Inc., 1994, pp. 61–69. [Online]. Available: http://portal.acm.org/citation.cfm?id=188508

[15] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker, "An empirical survey of linked data conformance," *Journal of Web Semantics (to appear)*, 2012.

[16] C. Fellbaum, *Organization of Verbs in a Semantic Net*. Kluwer, 1999, pp. 93–109.

[17] I. Niles and A. Pease, "Towards a standard upper ontology," *Proceedings of the international conference on Formal Ontology in Information Systems FOIS 01*, vol. 2001, pp. 2–9. [Online]. Available: http://portal.acm.org/citation.cfm?doid=505168.505170

[18] J. Conesa, V. Storey, and V. Sugumaran, "Experiences using the researchcyc upper level ontology," in *Natural Language Processing and Information Systems*, ser. Lecture Notes in Computer Science, Z. Kedad, N. Lammari, E. Mtais, F. Meziane, and Y. Rezgui, Eds. Springer Berlin / Heidelberg, 2007, vol. 4592, pp. 143–155.

[19] J. Rao and X. Su, "A survey of automated web service composition methods," *Semantic Web Services and Web Process Composition*, vol. 3387, no. 2, pp. 43–54, 2005. [Online]. Available: http://www.springerlink.com/index/4M6W37G0JFFK9BV4.pdf

[20] C. Garrod, A. Manjhi, A. Ailamaki, B. Maggs, T. Mowry, C. Olston, and A. Tomasic, "Scalable query result caching for web applications," *Management*, vol. 1, no. 1, p. 550561, 2008.

[21] H. Agt, G. Bauhoff, R.-D. Kutsche, and N. Milanovic, "Modeling and analyzing non-functional properties to support software integration," in *CAiSE Workshops*, ser. Lecture Notes in Business Information Processing, C. Salinesi and O. Pastor, Eds., vol. 83. Springer, 2011, pp. 149–163.

[22] C. V. S. Prazeres, C. A. C. Teixeira, and M. D. G. C. Pimentel, "Semantic web services discovery and composition: Paths along workflows," *2009 Seventh IEEE European Conference on Web Services*, pp. 58–65, 2009.

[23] M. O. Shafiq, R. Alhajj, J. G. Rokne, and I. Toma, "Lightweight semantics and bayesian classification: A hybrid technique for dynamic web service discovery," in *IRI*. IEEE Systems, Man, and Cybernetics Society, 2010, pp. 121–125.

[24] P. Plebani and B. Pernici, "Urbe: Web service retrieval based on similarity evaluation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 11, pp. 1629–1642, 2009.

[25] J. Cardoso, A. Barros, N. May, and U. Kylau, *Towards a Unified Service Description Language for the Internet of Services: Requirements and First Developments*. IEEE, 2010, p. 602609.

[26] M. Sabou, C. Wroe, C. Goble, and G. Mishne, "Learning domain ontologies for web service descriptions: an experiment in bioinformatics," in *Proceedings of the 14th international conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005, pp. 190–198. [Online]. Available: http://doi.acm.org/10.1145/1060745.1060776

[27] U. Chukmol, A.-N. Benharkat, and Y. Amghar, "Enhancing web service discovery by using collaborative tagging system," *2008 4th International Conference on Next Generation Web Services Practices*, pp. 54–59, Oct 2008.

[28] S. Kona, A. Bansal, G. Gupta, and T. D. Hite, "Web service discovery and composition using usdl," *The 8th IEEE International Conference on ECommerce Technology and The 3rd IEEE International Conference on Enterprise Computing ECommerce and EServices CECEEE06*, pp. 65–65, 2006.

A Generic Testing Framework for the Internet of Services

Senol Arikan, Aneta Kabzeva, Joachim Götze, Paul Müller
ICSY - Integrated Communication Systems
TU Kaiserslautern
Germany
{arikan, kabzeva, j_goetze, pmueller}@informatik.uni-kl.de

***Abstract*— A widespread approach of the design and development of heterogeneous distributed software systems is the use of an interacting group of services. This approach uses the concepts of Service-oriented architectures to realize a dynamic adaptive communication system among service provider, service consumer and service broker. Today, the Internet, as the largest heterogeneous distributed software system, uses the ideas of service orientation more and more, thus it is extended to the Internet of Services. In the Internet of Services, autonomous services can be deployed by different service providers on service platforms; thereby they are available via the Internet for a large number of service consumers. Service providers can change their service implementation at any time without notifying the service's consumers; therefore, no guarantees can be made about the adherence of these services to the specification the consumers expect. In order to ensure the compliance of the services to their specifications and provide a certain level of quality assurance for service consumers and platform providers, a service platform needs to provide comprehensive testing mechanisms which support the quality needs of all actors on the platform. In this paper, we propose a generic testing framework which can be used during design and run-time for the automated verification of distributed services.**

Keywords- SOA; Internet of Service; run-time testing; black-box testing; verification

I. INTRODUCTION

The quality of large software systems is one of the most important goals of software development. Balzert [8] defines software quality as the sum of the following characteristics: functionality, usability, reliability, performance, maintainability. All these characteristics define the degree to which a software product fulfills its functional and non-functional requirements.

One fundamental prerequisite for software quality is the software's robustness to possible faults. This can, for example, be achieved through early identification and correction of failures [12]. In general, a failure in a system means that a wanted behavior is not achievable (i.e., a behavior which does not conform to the requirements has occurred. In order to detect such a system behavior, different testing strategies can be performed. A fundamental classification divides testing strategies into black-box and white-box testing.

The complexity of software testing reaches a new level with the need to test heterogeneous distributed software systems. A widespread approach for the design and development of heterogeneous distributed software systems is the use of an interacting group of services. This concept is based on the architectural principle "separation of concerns", which focuses on one simple, well-known idea: a large problem is more effectively solved if it can be broken down into a set of smaller problems [11]. Service-oriented architectures (SOAs) solve complex concerns using service orientation. It is important to have loosely coupled services so that failures or changes in one service do not cause failures in other services. Service-oriented architectures realize a dynamic-adaptive communication system among service providers, service consumers, and service brokers. In the Internet of Services (IoS) [9], loosely coupled, reusable, autonomous services can be deployed by different service providers on service platforms. These services are distributed by the platform provider and, as such, are reachable over the Internet by a huge number of service consumers. Since services are offered by different service providers, no guarantees can be made about the functional adherence of these services' implementations to their previously defined specifications. Providers can change their service implementations and introduce new bugs at any time without notifying the service consumers [14]. Additionally, platform providers cannot be sure about the performance of their services. In order to test the services' compliance to their specifications and provide a certain level of quality assurance for service consumers and platform providers, service platforms need to provide comprehensive testing mechanisms which support the quality needs of all actors on the platform.

Unfortunately, platform providers typically do not have access to the services' source code as a standard service design principle, abstraction, is that no information about the internal realization of a service has to be published for other actors of the distributed system. Hence, the implementation of a service is unknown for a service consumer and platform provider. This fundamental characteristic along with the need for service providers and platform providers to ensure the quality of their services forces service testing to focus on black-box testing approaches in the Internet of Services. The service tester,

ISBN: 978-1-61208-200-4

independent of his role on the platform, should be able to invoke a service with a specific test case to check the response of the service. If the service output doesn't conform to the expected value, then the service does not meet the expected quality for that test case.

To ensure the conformity of a service implementation with its service specification during its complete lifetime, a service must be tested not only during development but also at run-time. We propose a generic testing framework which can be used for the automated verification of services during their complete life cycle. The proposed solution is based on black-box testing. An evaluation of the concept is provided by a prototype implementation of the framework in an existing SOA-based infrastructure. The framework takes as input user-defined valid test cases and generates test clients, which are executed to try out the desired service functionalities for suitable parameters. Analysis of the test results and notification of affected actors are also important requirements to address the challenges of a testing framework for the Internet of Services.

In order to address the quality requirements of all relevant stakeholders at any time in the service lifetime, the framework supports stress tests, scalability tests and parallel tests. To be able to support the testing of the potentially large number of resources offered on a service platform and the dynamic number of testing requests coming from consumers, the proposed solution considers asynchronous and synchronous communication models.

The paper is structured as follows. In the next section, we present some related work. In Section III, we explain the basics needed for understanding the work. Section IV describes the challenges for testing in the context of Internet of Services in more detail. Section V presents the architecture of the testing framework proposed as a solution addressing these challenges followed by an implementation approach. Section VI concludes the paper and discusses identified future work.

II. Related Work

In this section, we give an overview regarding the different solution proposals from other researchers for testing service-based distributed systems. The approach of Looker, Munro and Xu [1] concentrates on measurement techniques to test the robustness of Web services using *network level fault injection* to manipulate the expected parameters of a service at run-time. This approach has the disadvantage of requiring the service's source code in order to make required modifications. A service tester who does not have access to the service's implementation cannot use this approach to test a service. Our framework uses a black-box technique, thus it enables testing of a service for each stakeholder involved in the life cycle of a service. Frantzen, Tretmans, and Vries [2] apply a *model-based testing technique* to experiment with a Web service, which aims at either finding faults or gaining confidence in the service. Model-based techniques have been developed for reactive systems. In order to apply techniques for MBT (Model-Based Testing) of reactive systems in SOA-based systems, some additional requirements must be satisfied. These additional requirements can increase the complexity of the realization of the proposed approach.

Most of the solutions for testing of service-based distributed software systems have experimented with SOAP-RPC based Web Services. Chakrabarti and Kumar [3] have developed a black-box approach for testing RESTful Web Services which uses a test specification language for better automation in test execution. This approach is limited to testing RESTfull services.

To the best of our knowledge, the only works which address issues close to ours are [4] and [5]. Martin, Basu, Xie [4] presents a unit testing framework for Web services based on JUnit. This framework uses the test generation tool JCrasher in order to generate corresponding JUnit tests. WS-TAXI [5] is a WSDL-based testing tool for Web Services, which is obtained by soapUI [6], an industrial testing tool, and TAXI [7], which automatically generates XML-based test cases from a corresponding XML schema. This framework is based on the idea of automatic generation of SOAP envelopes by using data instances from WSDL descriptions, which are used for service invocation. Our framework does not use any external tools for the generation of test cases. With only minimal amount of input data, which are given by service testers in XML-format, and with use of WSDL descriptions, the test clients will be automatically generated and executed at run time. We concentrate on the quality of the testing process and developing an efficient and dependable framework, which is highly performant and supports service testers during the whole testing process. Finally, the testers will be notified about analyzed test results. In next section, we go into details and present some basic terms and strategies for testing.

III. Testing Basics

In this section we will first define some of the terms that are commonly used when discussing testing. Then we will discuss the details of the two basic testing approaches - white-box and black-box testing, which were mentioned above.

A. Error, Fault, and Failure

There is considerable confusion regarding definitions of error, fault, and failure in the literature. We use the definitions from Jalote for these terms [14]. The term error is used to refer to any activity of a programmer which results in software containing a defect or fault. A fault is a condition that causes a system to fail in performing its required function. A failure is the inability of a system or component to perform a required function according to its specification.

B. Validation and Verification

In general, there are two important evaluation methods to check software against its specification: *verification* and *validation*. As defined by the IEEE [16], *verification* is a process of evaluating a system or component to determine whether the products of a given development phase satisfy the conditions imposed at the start of that phase; *validation* is the process of evaluating a system or component during or at the end of the development process to determine whether it satisfies specified requirements. Using these definitions, *validation* is a process to demonstrate that the software implements each of the functional requirements correctly and completely; *verification* is the process to ensure the software product of a given phase fully implements the inputs to that phase. The framework proposed in this work can be used in any service life cycle stage and therefore supports both the verification and validation of services.

C. *Software testing approaches*

Software testing approaches traditionally divide into black-box and white-box testing.

White-box testing approaches consider the internal data flow and logic of the system under test. This approach is also known as glass-box testing or structural testing. The internal working of the software is visible for the tester. Because the implementation of the software product is known, white-box-testing enables a tester to design test cases that exercise the independent paths within a module or unit, check logical decisions on both their true and false side, execute loops at their boundaries and test the validation of the internal data structure [13]. White-box testing gives a tester a certain amount of control during the testing process. If a fault is detected, the tester knows which lines of code to look at based on the corresponding test case. Because of this control, defining and removing faults in the tested object is more economical and successful than with other testing approaches. The internal data and logic flow of a service is known only to the service developer in the context of Internet of Services. A service provider is not necessarily a service developer and therefore will have no knowledge of the details needed for specifying white-box test cases. This is also the case for consumers of these services.

Black-box testing approaches, also called behavioral testing, consider the tested system as a whole and ignore internal structure details. In contrast to white-box testing, black-box testing is usually used when the implementation of the software is not known to the tester. Black-box testing uses the functional requirements and specifications of the software to define test cases that should fully exercise all the functional requirements [14]. These resources are available for platform actors in IoS: each service has to provide a specification of its functionality which an interested actor can use to define a desirable test case. After generation of the test cases from a specification, some valid and invalid input data are provided for test execution, and then the testing method calls the corresponding software to verify whether the test results are compatible with the expected outputs. Black-box testing terminates when all test cases are executed. According to Pressman [13], black-box testing techniques are applied to find errors in the following categories: incorrect or missing functions, interface errors, errors in data structures or external database access, behavior or performance errors, and initialization and termination errors. An important disadvantage of black-box testing is that it does not help in finding the reason of the failure. Testing the code (implementation) quality is not possible.

Another well known problem of black-box testing is the selection of test cases. In order to deal with this problem, some black-box strategies were defined; they differ according to test case selection criteria [12]. The first strategy is Equivalence Class Partitioning. The idea behind this strategy is to reduce the complexity of selecting test cases by dividing the set of all possible inputs for a function into a set of equivalence classes so that if any test in an equivalence class succeeds, then every test in that class will succeed [14]. Experience shows that faults often occur on the boundaries of equivalent classes [12]. Boundary Value Analysis is based on the experiences gained through Equivalence Class Partitioning and selects test cases which lie on the boundaries of equivalence classes. The goal is to reach a maximal number of tests with as few test cases as possible. Thus the complexity of the testing process can be reduced. Another strategy for black-box testing is Cause-Effect Graphing [14]. This strategy attempts to combine inputs from different input classes through the use of the Boolean operators “and”, “or”, and “not” in order to exercise some special test cases. The disadvantage of this approach is that it can result in a large number of test cases, many of which will not be useful for detecting new faults.

The prototype implementation of the proposed testing framework uses a randomized algorithm which gets random service relevant test cases from a database and executes them. As part of our future work we will specify algorithms based on the Boundary Value Analysis strategy.

In conclusion, black-box testing is not an alternative to white-box techniques. It can be considered as a complementary approach that returns a different class of errors than white-box testing [13]. It means we can also combine both strategies to test a software system, if the corresponding requirements (i.e., availability of source code) can be met.

IV. INTERNET OF SERVICES

In this section we will first introduce concept Internet of Services and then we will define some challenges which should be considered by the testing framework.

A. *Introduction*

With the adoption of the SOA paradigm for the design of distributed business processes [11] and the introduction of Cloud infrastructures that allow on-demand delivery of IT resources [21], the offering, discovery, and usage of technical services over the Internet is not a vision anymore, but a fact. Considering services as tradable goods over the Internet is the main concern of the Internet of Services community [9]. Yet, the opportunity to offer services reachable by a wide range of potential customers on platforms provided by third parties gives rise to some new roles. Each of these roles has its own requirements for the quality of service provided by the services offered on such a platform. Thus, a generic testing framework, used as a major quality control mechanism, should be able to address some of the new role-specific requirements.

Besides the typical software engineering roles of provider and consumer, the SOA paradigm introduces the role of a service broker, which serves as an intermediary between the provider and consumer [11]. In the Internet of Services, the three classical SOA roles are considered insufficient [18][19]. Additionally, the platform provider and service developer roles have to be considered.

In the Internet of Services, the service broker role is taken by the platform controlling the life-cycle of the services it offers [18]. A platform does not only manage a catalogue of offered services and their descriptions, but has also takes care of platform-wide security and quality standards. Stakeholders carrying these responsibilities in the Internet of Services are referred to as platform providers. The platform provider is responsible for providing qualitative infrastructure whose customers are the service providers.

Compared to traditional mainframe applications where the operating organization is normally also the supplier of the software, the Internet of Services makes a distinction between service developer and service provider [19]. Nevertheless, these two roles are not mutually exclusive. A service developer concentrates only on writing the executable code behind a service. This is the only role that has knowledge of the internal logic and data flow within the code. The rest of the stakeholders only have access to the service through its interface description. A service developer has to guarantee the quality of the code only against the service provider. Service providers are responsible for the deployment of services on a platform and the specification of service level agreements (SLAs) [22]. They provide services that offer some value to the service consumers and use the resources of the platform to communicate with their customers. A service provider is accountable for granting the quality of service specified in the SLAs and for compensations in case of violations.

The service consumer uses the platform to find one or more services which can fulfill his needs. The product which is of interest for the service consumer is the real-world effect provided by the functionality of a service. Once a suitable service is identified, a contract has to be negotiated between the consumer and the provider of the service [20]. Since a selected service will probably be integrated in the consumer's operational environment, the consumer has to be given the possibility to test if the service quality still fits the requirements of his own environment at any time.

In addition to the extended number of roles in the Internet of Services, the dynamic organization of service-based distributed systems also introduces some challenges to the execution of tests in such an environment. The changing number of stakeholders acting on a Cloud platform may lead to a large number of test cases that should be covered by the platform testing framework. Since some of the stakeholder roles, like the service developer and the service provider, are interested in design-time tests, and all roles have to be able to check the compliance of the resources to the negotiated contracts at any time, tests should be executable at both design and run time. Some test cases would be executable on demand (i.e., after changes or failure corrections). Others, like tests checking the compliance with SLA terms, should be executable on a regular basis. When quality violations are discovered in the testing process, the testing framework should be able to send the right information to all affected stakeholders; which requires the integration of a notification mechanism within the framework.

In the following section, we present a list of challenges for testing SOA-based distributed software systems from an IoS perspective.

B. *List of challenges*

Considering the relationships, responsibilities, and organization of a service platform in an IoS environment, we identified the following challenges that should be addressed by a testing framework:

- *Large number of test cases*: the number of stakeholders interacting on the platform is variable. Any number of users can join the platform; any number of services can be deployed on the platform. As a consequence, the platform must be able to provide for the execution of the continuously growing number of test cases by making scaling the framework a core part of the implementation.
- *Lack of knowledge on service structure:* for all stakeholders except for service developers, services are only known through their interfaces, the service implementation and structure are intentionally hidden. This makes white-box testing impossible and forces black-box testing.
- *Service life-cycle:* once deployed a service should be always available and cannot be taken offline for maintenance. Thus it is important to provide testing support during service development as well as during service run-time.
- *Different responsibilities*: depending on their role on the platform, different stakeholders have different responsibilities, as explained in the previous section. A testing framework for the Internet of Services should be able to support different kinds of testing in order to cover all role-specific needs.

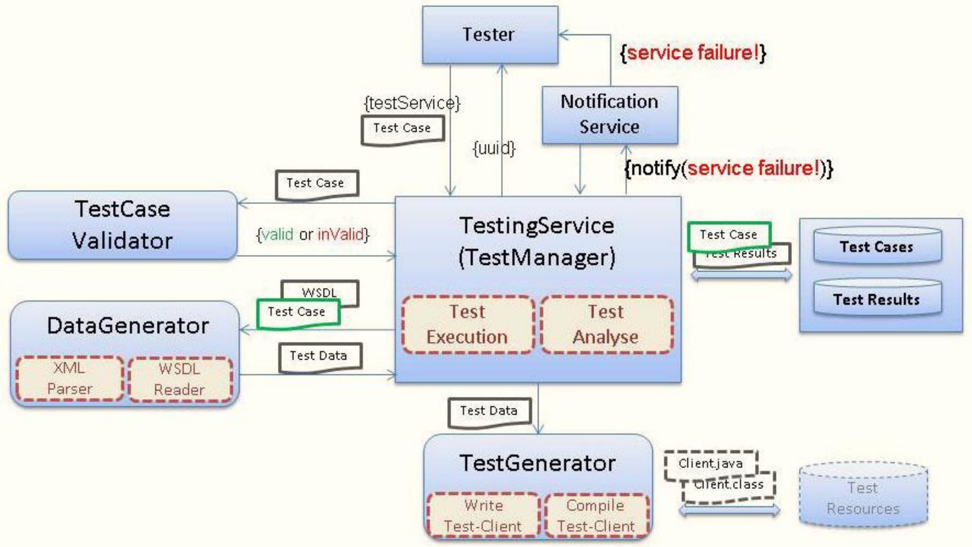


Figure 1. Architectural Layout of the Venice Testing Framework.

- *Different requirements:* different stakeholders have different perspectives on the platform, requiring support for a variety of use cases. On-demand testing and periodic testing should be supported in order to address the different needs of the service tester for the separate test cases.
- *Large amount of data*: the execution of a large number of test cases will produce a large quantity of data. The platform must be able to provide storage and analysis for a large quantity of test results.
- *Lack of trust:* access to testing data and results is a trust issues in the open environment of the Internet of Services. Stakeholders should be granted proper handling of the data they provide for testing purposes. Security mechanisms regulating the access to this data should be assured.
- *Lack of evolution control:* a service provider can change the functionalities of a running service at any time. This can result in an unexpected change for some of its users. In order to prevent this situation, service users affected by a change must be informed about service modifications.
- *Dynamicity*: the dynamic character of SOAs enables new services to be deployed on the platform, existing services to change, or removal of unused or defective services from the platform. The framework must automatically perform acceptance testing [8] on deployment of new services to ensure the quality of the resources offered on the platform. Regression tests [12] must be executed on every change of existing services to ensure compliance with existing SLAs and contract terms. Deactivation of test cases for deleted services should also happen automatically in order to prevent unnecessary resources usage.

V. Solution Architecture

The proposed testing framework enables Web service developers and other stakeholders to automatically and fully test services during development and run time. If an error occurs during the actual service execution (e.g., a service cannot be reached, or its output does not correspond to expected values) all participants of the testing process will be notified about this error.

For the usage of the testing framework two use-cases can be defined. A service developer can use the framework to check the functionalities of a service during development time. The framework also can be used to test the services at run-time. This use case scenario is useful especially for platform providers and service consumers. A monitoring service can navigate the testing framework to execute test cases on the basis of a predefined test schedule. Services can be tested on-demand or periodically.

The architecture of the framework is shown in Figure 1. The framework is composed of several components - *TestManager*, *TestCaseValidator*, *DataGenerator*, *TestGenerator*, a *database*, and a *repository* for test resources – which, in combination, execute the framework functionalities. It uses also the Notification Service of the Venice Framework [10] to keep all participants informed of test results.

A. Testing life-cycle

The framework supports all four phases of the defined testing life-cycle: test specification, test organization, test execution, test analysis. In the following, we describe functionalities of the framework components on the basis of these life-cycle phases.

1) Test specification

In order to execute a functional test, the testing framework needs some input data. This information should be defined by a service developer in XML. Our framework offers a XML schema to support the tester during the description of the test cases.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

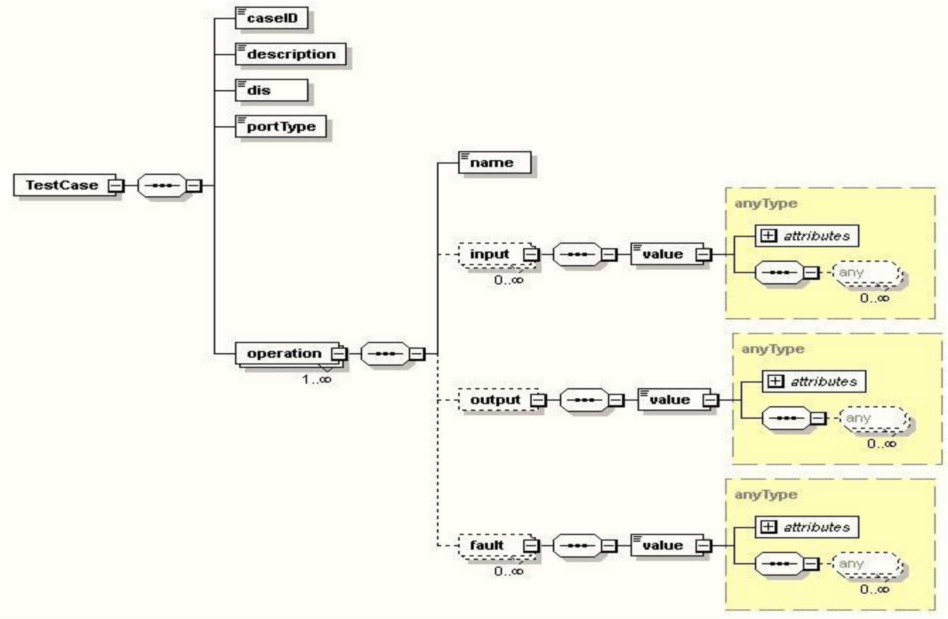


Figure 2. XML-Schema to define test cases.

Figure 2 presents the test case XML schema. A test case must have a unique name (*caseID*). The corresponding test case will be stored in the database with this name. Further important information in test cases are the Domain Information Service (dis) and port type (portType) fields. The dis provides meta-data for the service domain and enables service interaction in the Venice environment [10]. The portType defines an abstract name for a set of operations and messages. A test case has to define which operations will be tested along with its input and expected output. Each operation can also have a fault element, which should demonstrate a service call for an invalid input.

The framework offers operations to add test cases into a database, to get them and to delete them. Before storing into the database, the test cases have to be checked for their validity by the component *TestCaseValidator* (see Figure 1). Only valid test cases will be used.

2) Test organization

Different options for testing are offered to the tester. The tester can test all the functionalities of a service, meaning that all the port types implemented by the service will be tested. Platform providers can use this operation before the deployment of the new service on the platform for acceptance testing. Service consumer can test the entire functionality of a service through this operation.

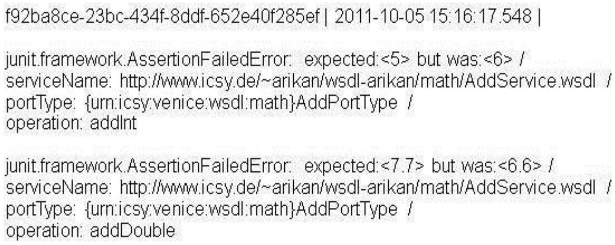
The tester can also test all the services which implement a certain port type. This operation is useful for a platform provider to perform automated tests for the complete platform. This also allows service consumers and platform provider to run performance tests or stress tests. Another useful operation is for creating a new test, which a service consumer can use to define a test case and then execute it.

3) Test execution

After a successfully validation of a test case against the test case schema (passing a syntax check), the test case will be parsed by the *DataGenerator* component of the framework, which also uses the WSDL description of the service to get more data. All test data (from the test case and the corresponding WSDL) will be encapsulated in a *TestCase* object and sent to the *TestGenerator*. The *TestGenerator* generates and compile a JUnit-based Java test. The resources are stored in a repository, which is created at the beginning of the testing process and deleted at the end of the testing process. After compilation, the newly generated test case will be executed.

4) Test analysis

If an exception is captured during the execution of the tests, this will be stored in a *TestResult* object. All test results will be written into the *TestResults* database (see Figure 3). These test results can be retrieved with the *getTestResult* operation of the Testing Service.



```
f92ba8ce-23bc-434f-8ddf-652e40f285ef | 2011-10-05 15:16:17.548 |

junit.framework.AssertionFailedError:  expected:<5> but was:<6> /
serviceName: http://www.icsy.de/~arikan/wsdl-arikan/math/AddService.wsdl  /
portType: {urn:icsy:venice:wsdl:math}AddPortType  /
operation: addInt

junit.framework.AssertionFailedError:  expected:<7.7> but was:<6.6> /
serviceName: http://www.icsy.de/~arikan/wsdl-arikan/math/AddService.wsdl  /
portType: {urn:icsy:venice:wsdl:math}AddPortType  /
operation: addDouble
```

Figure 3. An example to present the testing results for Add-Service.

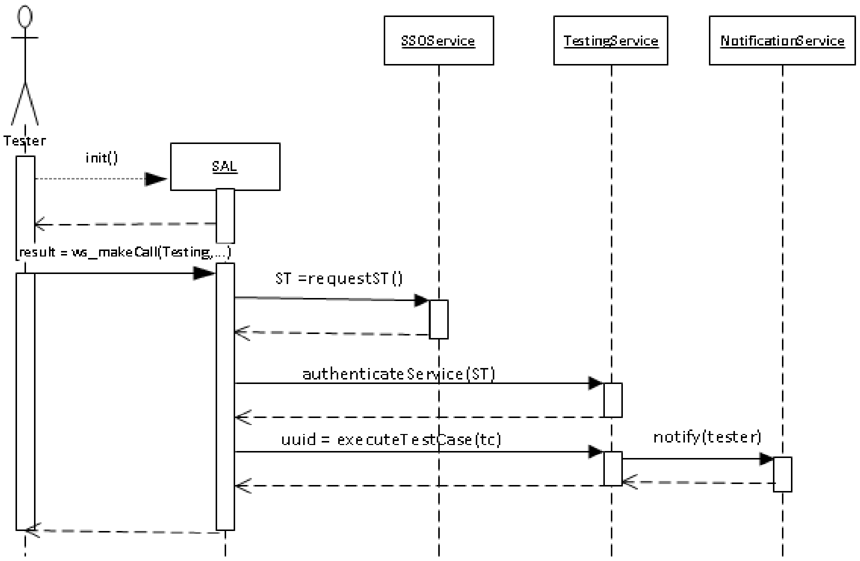


Figure 4. Use case diagram to demonstrate using of the Testing Service

B. *Implementation Prototype*

We implemented a prototype of the testing framework for the Venice (see Figure 4) platform. Venice is a SOA-based framework for building secure and dependable distributed applications; it supports service developers during developing, deployment, maintenance, and usage of Web services. Different service providers can use the Venice infrastructure to offer their services for service consumers. Figure 4 shows how the testing framework is used in the Venice environment. In order to use the testing framework, the tester first needs to initialize the *Service Abstraction Layer (SAL)* of Venice. The SAL accesses to additional functionalities like authentication and authorization, which are provided by the Venice Single Sign-On Service (SSO Service). To use the testing service, service consumers have to authenticate one time to the *SSOService*, which returns a service token (ST). The ST contains the authorization information that allows the user to prove his identity and to prove his right to access the testing service. All necessary operation invocations are made transparently for the service consumer. The next step is calling the desired operation of the testing service. After the testing process is finished, the tested service returns a unique *uuid*, which is used to request test results from the database. Service consumers will be informed of the test results through the notification service provided by Venice. Finally, test results are fetched from the database.

The testing framework is implemented in Java and uses the *JUnit* libraries. Figure 5 shows the implemented classes of the framework and their relationships.

The Testing class uses the *InputGenerator* to get input data as a *TestCase* object. The *InputGenerator* uses *DOMParser* to parse a test case, which was created by service developer in XML. The *DOMParser* class reads XML files and generates the corresponding input, output and fault objects, which will be added to a *TestCase* object. A *TestCase* object will be given back to the Testing class. It uses the *WriteUniTest* class to generate java test classes. These will be compiled, and then executed by *MyTestSuite*.

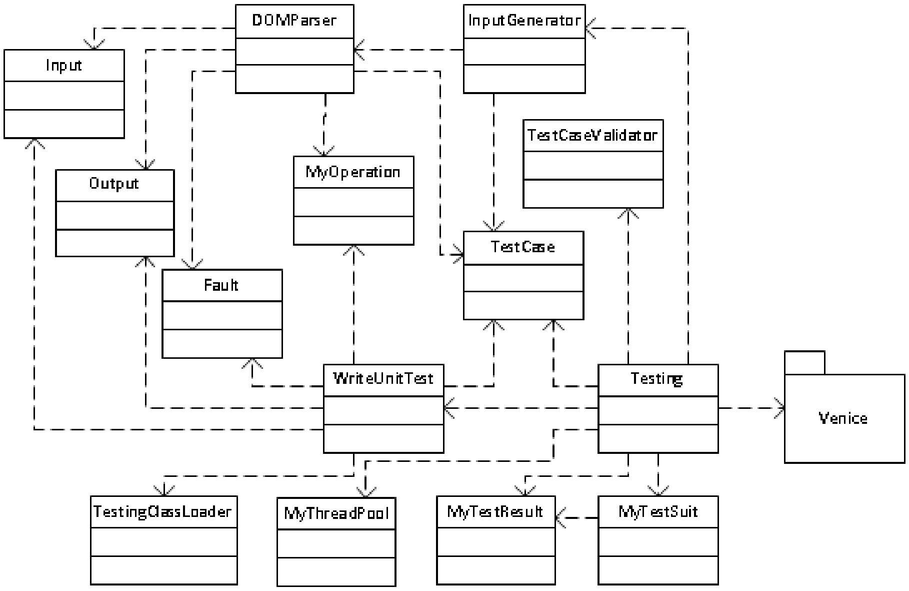


Figure 5. Static structure of the Testing Framework.

Test results will be added to a *MyTestResult* object and stored in the database. The clients will be notified through Venice's notification system. To perform incoming tasks more efficiently, we implemented a thread pool. Tests are temporarily stored in the *IncomingRequests* queue and are executed by worker threads in the thread pool.

VI. Conclusion and Future Work

In order to meet new quality requirements in software development, software testing has been researched for several years. With the application of SOA as a concept for development of distributed services on the internet –Internet of Services - new challenges for testing infrastructures were defined. In order to satisfy these challenges, we designed and implemented a generic testing framework. Our proposed framework is based on black-box testing and supports the whole testing life-cycle; from generation and checking of the test cases to compiling and execution of test cases.

This paper presented a generic testing solution and its prototype implementation for testing of IoS service platforms. In future, the functionality of the framework will be extended; we plan to enable the test result analysis and present statistics for executed test cases. Furthermore, in order to provide better performance and scalability, an asynchronous communication pattern will be implemented and integrated into the prototype. A graphical user interface is planned in order to increase usability.

VII. Acknowledgements

This work was funded by the German Ministry for Education and Research (BMBF project "iGreen" 01IA08005G). The authors are responsible for the content.

References

[1] N. Looker, M. Munro, and J. Xu, " Testing Web Services," the 16th IFIP International Conference On Testing of Communicating Systems, Oxford, 2004, unpublished.

[2] L. Frantzen, J. Tretmans, and R. Vries, "Towards Model-Based Testing of Web Services," Inter. Workshop on WS. - Modeling and Testing (WS-MaTe2006), Palermo, 2006, pp. 67-82.

[3] S. K. Chakrabarti, and P. Kumar, "Test-the-Rest: An Approach to Testing RESTful Web-Services," IEEE COMPUTATIONWORLD'09, Athens, 2009, pp. 302 – 308.

[4] E. Martin, S. Basu, and T. Xie, "Automated Robustness Testing of Web Services," Proceedings of the 4th International Workshop on SOA and Web Services Best Practices, Oct. 23, Portland, Oregon, USA., 2006, pp. 114-129.

[5] C. Bartolini, A. Bertolino, and E. Marchetti, "WS-TAXI: a WSDL-based testing tool for Web Services," icst, International Conference on Software Testing Verification and Validation, Colorado, 2009, pp. 326-335.

[6] soapUI, http://www.soapui.org, April 2012

[7] A. Bertolino, J. Gao, and E. Marchetti, "Automatic test data generation for XML Schema based partition testing, " IEEE Automation of Software Test, Minneapolis, 2007, pp. 4-11.

[8] H. Balzert, Lehrbuch der Software-Technik, Spektrum Akad. Verl., 1998, pp. 257.

[9] TEXO, Business Webs in the Internet of the Services, url: http://www.internet-of-services.com/index.php?id=276&L=0 , April 2012

[10] The Venice Service Grid, url: http://www.v-grid.info/html/pdf/The%20Venice%20Service%20Grid.pdf, April 2012

[11] T. Erl, Service-Oriented Architecture, Concept, Technology, and Design, Prentice Hall PTR, March 2009, pp. 290-291

[12] P. Liggesmeyer, Software Qualität – Testen, Analysieren und Verifizieren von Software, Spektrum Akad. Verl., Heidelberg, 2002.

[13] R. Pressman, Software Engineering: A Practitioner's Approach, McGraw Hill, Boston, 2001.

[14] P. Jalote, An Integrated Approach to Software Engineering, Springer Verl., 1997.

[15] S. Arikan, M. Hillenbrand, and P. Müller, "A Runtime Testing Framework for Web Services", 36th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA'10), Lille, France, 2010.

[16] IEEE Standard Glossary of Software Engineering Terminology, IEEE std 610.12-1990, September 1990, url: http://web.ecs.baylor.edu/faculty/grabow/Fall2011/csi3374/secure/Standards/IEEE610.12.pdf, April 2012

[17] C. Haubelt, J. Teich, Digitale Harware/Software-Systeme - Spezifikation und Verifikation, Springer Verl., 2010, pp.95-111.

[18] A. Kabzeva, M. Hillenbrand, P. Müller, and R. Steinmetz, "Towards an Architecture for the Internet of Services", 35th EUROMICRO Conference on Software Engineering and Advanced Applications (SEAA'09), Patras, Greece, 2009.

[19] C. Janiesch, R. Ruggaber, and Y. Sure, "Eine Infrastruktur für das Internet der Dienste", HMD - Praxis der Wirtschaftsinformatik, 45(261):71–79, June 2008.

[20] J. Spillner, M. Winkler, S. Reichert, J. Cardoso, and A. Schill., "Distributed Contracting and Monitoring in the Internet of Services", Ninth IFIP WG 6.1 International Conference on Distributed Applications and Interoperable Systems (DAIS 2009), vol. 5523 of Lecture Notes in Computer Science, pp. 129–142. Springer-Verlag, Berlin Heidelberg,2009.

[21] B. Sosinsky, Cloud Computing Bible, Wiley Publishing, Inc., 2011.

[22] OASIS, SOA-EERP Business Service Level Agreement Version 1.0, Commetee Draft 03, January 2010, url: http://docs.oasis-open.org/soa-eerp/sla/v1.0/SOA-EERP-BSLA-spec-cd03.pdf, April 2012

Industrial Automation Software: Using the Web as a Design Guide

Dirk van der Linden, Georg Neugschwandtner
Electromechanics Research Group
Artesis University College of Antwerp
Antwerp, Belgium
dirk.vanderlinden, georg.neugschwandtner@artesis.be

Herwig Mannaert
Department of Management Information Systems
University of Antwerp
Antwerp, Belgium
herwig.mannaert@ua.ac.be

***Abstract*—When looking the World Wide Web (and the Internet at large) as one giant application, we observe certain desirable properties that would also be welcome, but cannot be taken for granted, in industrial automation systems. In particular, subtasks that are unrelated from a functional point of view (such as Web sites) are usually well-separated from each other. Web services can play a key role in bringing separation of concerns to automation systems. The paper introduces a relevant standard (OPC UA). It also presents the Normalized Systems theory as a structured way of ensuring separation of concerns which is applicable to a wide range of application domains, from the Web to programs within automation controllers.**

Keywords-OPC UA; Normalized Systems; Evolvability; Industrial Automation.

I. Introduction

Already before the Internet was adopted worldwide, and, in parallel, IP-based networks obtained their dominant role in the office world (no matter in which sector of the economy) [1], some authors have tried to figure out whether the Internet could ‘crash’ or not. Ted Lewis answered the question ‘Is it possible for the Internet to overload and blow a fuse’ with ‘probably’ [2], following a (simplified) comparison of the Internet with an ‘infinite bus’, which can indeed become overloaded – at least partly. Consequently, the threat has to do with a potential lack of hardware resources, or even physical unavailability in case of, e.g., a fire in one or more buildings. However, a crash of the entire Internet is considered improbable [3]. Still, this only applies to the entire Internet as a system; parts of this system, in particular individual application servers, certainly can crash. To prevent services becoming unavailable due to such *object* crashes, authors for example consider ways to implement highly-available distributed World Wide Web (WWW) servers [4].

Considering this, the Internet appears to be a *stable* system. In the literature, the meaning of *stability* varies [5]. For a more formal specification of this assumption for our purposes, we shall consider the generic concept of stability as defined in the fields of signal processing, systems theory, and control theory, BIBO (Bounded Input Bounded Output) stability. In these fields, (BIBO) stability is considered one of the most fundamental properties of a system [6], [7]. It implies that a bounded input function should result in bounded output values, even as $t \to \infty$ (with t representing time).

When we call the Internet a *system* in this sense, we focus on the fact that objects, services, clients and servers are being continuously added to it, updated and removed from it. We consider these changes as inputs to the (Internet) system, and consider the impact of those changes as outputs. The Internet apparently copes well with these continuous, worldwide changes. Their consequences are bounded: An object crash does not affect the robustness of the system as a whole. Also, the effort for making changes does not rise as the system evolves: For example, creating a new web site is not harder today than last year (given that the new site should provide the same functionality and given the same tools and availability of other resources). Of course, adding a more complex web site requires more effort than adding a simple one; but the effort required does not depend on the size of the web at large. The effort only depends on the size of the change itself.

For software systems in general, however, this is not the case. Industrial automation control systems are no exception [8]. Stability or long-term maintainability is a challenge. It is a desired characteristic, which is hard to control [5]. The issue of evolvability is widely known, and was already specified in the form of a statement back in 1980 by Manny Lehman: ‘As an evolving program is continually changed, its complexity, reflecting deteriorating structure, increases unless work is done to maintain or reduce it’ [9].

Thus, the Internet appears to be stable, while the software systems, which make part of it, are not. This paper presents approaches towards improving this situation, focusing on separation of concerns in automation systems by applying web based technology and appropriate formal guidelines. Section II outlines the desirable properties we find in the Internet as a system and how automation systems are different in contrast. Section III introduces the OPC Unified Architecture standard as a way of applying web based technology to automation systems. Section IV summarizes the main theorems of the Normalized Systems theory, which provides formal guidelines to building stable software systems independent of the application domain, and Section V

ISBN: 978-1-61208-200-4

discusses how to migrate industrial automation systems to normalized systems. In Section VI, we conclude and present suggestions for future research.

II. The Web as a model for separation between automation software components

When looking at the world wide web, we notice various desirable properties. We already mentioned the example of web sites being independent of other web sites (as long as they do not depend on each other for content) regarding the effect of crashes as well as the effort required for launching a new website. Also, a web browser is not going to crash when a server does not respond in time. The Internet is robust against changes, *without* the need for a specific maintenance effort as stated by Lehman's law [9]. We are seeing a huge, stable system where essential concerns are well separated. Broadly speaking, each web site could be considered such a concern.

While our examples for desirable properties found in the world wide web are probably so familiar that they do not seem to be a special achievement at first sight, it is worth remembering that these properties cannot simply be taken for granted. As a case in point, the situation is significantly different with industrial automation systems.

Industrial Programmable Logic Controllers (PLCs) are usually programmed in one of the languages of the IEC 61131-3 standard [10]. An IEC 61131-3 Program Organization Unit (POU) contains code, which can be a Function, a Function Block, or a Program. The code can be written in any one of these languages, depending on which of them is most appropriate for the specific application.

Often a PLC controlled factory plant starts out from a basic solution and is extended over time. For such a basic solution in a starting SME (Small or Medium-sized Enterprise) business, one single PLC might be enough to implement production control (or a self-contained subpart). If the business is successful and the production capacity has to be expanded, the resource limit of the single PLC will be reached at some point. Typically, a second one is then added, which will result in a number of couplings between these PLCs. In case engineers are focusing on functional requirements only and neglect non-functional quality properties (like evolvability and stability), there is a risk that these couplings could be so-called undesired couplings [11]. These will cause combinatorial effects, resulting in an increased impact of changes as the system size increases. Engineering an evolving system based on functional requirements only – and thus neglecting desirable maintenance activities with respect to the software – typically results in a system with a low amount of separable concerns. Without separation of concerns, it is likely necessary to shutdown the entire system if one of the many PLCs needs to be replaced; it could even be necessary to re-engineer the entire system. All this is due to the negative effect of coupling between POUs – in contrast to the loose coupling that we observe to be in place between web sites.

One of the very likely sources of undesired coupling is that popular communication systems between PLCs are based on global variables, shared memory, or shared I/O (Input/Output) addresses (e.g., fieldbus systems). The *existence* of these systems is not necessarily a disadvantage, but the *way of use* of these mechanisms can lead to invisible, hidden dependencies (also referred to as 'common coupling' in this case [11]), resulting in combinatorial effects.

Of course, it is possible to create IEC 61131-3 programs without causing common coupling (e.g., by carefully documenting the use of all global variables). However, instead of hoping that developers use these communication capabilities without causing common coupling (which would at least require thorough education), it would be preferable to have concepts in place that actually remove the possibility of common coupling [12].

Another relevant aspect in this context is the separation of states. A simple example where this is achieved on the world wide web would be the separation between web browser and server in the example at the beginning of this section. On the other hand, in industrial automation, some PLCs contain an integrated diagnostic system to handle fieldbus failures that can actually cause the PLC to shutdown if handling code is not properly implemented. In other words, instead of just notifying the PLC system of a fieldbus failure, a fieldbus system failure can be propagated to cause further system failures.

III. OPC Unified Architecture

Web services cleanly separate software components from each other. They enable self-describing, modular applications to be published, located, and invoked across the Web. Software modules collaborating via web services make parameters and arguments explicit through the module's interface only. They allow modules to be loosely coupled. Hence, combinatorial effects cannot propagate. In other words, web services isolate web based applications from each other. Therefore, we think that web services are a very helpful means toward avoiding common coupling as outlined in the previous section. We think that the structure they bring to systems by way of their design can contribute significantly to system stability.

The use of service oriented architecture and web services in manufacturing systems has been previously studied in depth by academic research groups [13], [14]. However, the lowest level of control (also called the shop floor layer) is characterized by a great heterogeneity of systems and special fieldbus communication solutions. Unless they support a common communication standard, the effort to integrate these shop floor systems into a web services based solution is prohibitive in industrial practice. Here, the OPC specifications could play a major enabling role.

The OPC Foundation set out to enable interoperability between automation equipment and software of various vendors. The first and still most successful specification, called OPC Data Access, was designed as an interface to communication drivers, allowing standardized read and write access to real-time data in automation devices. The major use case are HMI and SCADA systems accessing data from different types of automation devices. The first OPC specification family (released in the nineties) was based on the DCOM (Distributed Component Object Model) technology by Microsoft.

OPC has become the de facto standard for industrial integration and process information sharing [15]. By now, over 20,000 products are offered by more than 3,500 vendors. Kalogeras et al. share the view that OPC is highly accepted in industry [16]. Millions of installed OPC based products are used in production, process industry, building automation, and many other fields of application around the world [17]. However, when the Internet gained widespread adoption, the DCOM technology caused certain limitations.

To address this, the OPC UA specifications were released in 2006, aligning OPC with the principles of service oriented architecture. OPC UA is considered one of the most promising incarnations of WS technology for automation [18], [17], [19]. Its design takes into account that the field of application for industrial communication differs from regular IT communication: embedded automation devices provide another environment for Web-based communication than standard PCs.

OPC UA makes it feasible to provide a web services based OPC UA interface directly on automation controllers, with integrated protocol security. OPC UA can thus be used to encapsulate PLC programs much in the way that HTTP and HTML ‘encapsulate’ the implementation details of a web server. For this purpose, an OPC UA companion specification for representing IEC 61131-3 code using the information modeling mechanisms of OPC UA is available.

While separation of concerns can also be achieved by using proprietary, custom-made web services, using OPC UA instead has additional benefits. The most important one is that, being a standardized interface, it enables interoperability between automation systems by different vendors. Given the fact that web services interfaces will usually be provided by automation system vendors, the use of non-standard interfaces makes considerable integration efforts necessary in a multi-vendor system (Figure 1). Also, automation system integrators typically do not have the skills to develop custom-made webservices. Rather, they are trained to write IEC 61131-3 code and configure commercial SCADA products, often with a C-like scripting language. The OPC UA specifications are an excellent way to stimulate automation product vendors to provide standardized interfaces, which can be configured by automation system integrators in webbased or web-enabled automation software components.

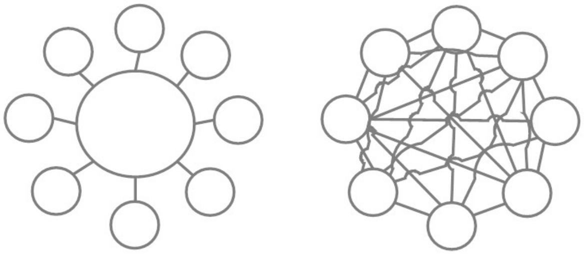


Figure 1. Left: OPC UA based communication; Right: Communication based on custom webservices

IV. Normalized Systems

The value of applying separation of concerns throughout a system (in our case, an industrial automation system) is apparent. Thanks to OPC UA, web services can be more easily applied to achieve loose coupling between PLCs, as well as between PLCs and SCADA systems. Still, unwanted couplings can also exist between POUs within a PLC. Moreover, applying web services technology is only one step towards achieving loose coupling; interfaces still have to be designed carefully to avoid unwanted functional dependencies. While comprehensive informal guidelines are available to educate software developers about good design, few formal contributions exist.

The Normalized Systems theory has recently introduced an approach to attain evolvable modularity in software, starting from a constructive point of view. Its authors state that probably all necessary knowledge is available to build stable software systems, but it seems to be hard to apply this knowledge [20]. The reason why it is very challenging to build stable software systems is not due to a lack of knowledge, but because this knowledge mainly takes the form of developers’ individual experience. The Normalized Systems theory attempts to capture this knowledge in a succinct and formal way.

In Normalized Systems theory, instead of taking the functional requirements as the only starting point, elementary software constructs are defined as basic building blocks. Hence, the implementation of software can be seen as the transformation of functional requirements into software primitives. We can represent this implementation transformation γ as

$$\mathcal{S} = \gamma(\mathcal{R}) \quad [21].$$

In this formula, S represents the set of elementary software constructs, and R stands for the functional requirements.

In order to obtain evolvable modularity, the Normalized Systems theory states that this transformation should exhibit systems-theoretical stability. This means that a bounded input function (i.e., a bounded set of requirement changes) shall result in a bounded set of output values (i.e., a bounded

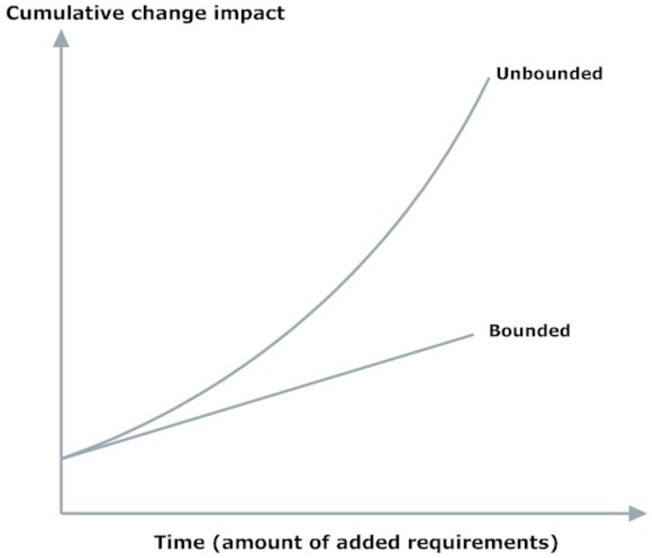


Figure 2. Cumulative impact of changes over time [24]

impact or effort), even if an unlimited systems evolution is assumed. The impact of a change shall only depend on the nature of the change itself. Conversely, changes causing impacts that are dependent on the size of the system are termed *combinatorial* effects; they must be eliminated from the system in order to attain stability. Stability in this context amounts to a linear relationship between the amount of requirements on the system (which is increasing as the system evolves over time) and the effort required to implement all these requirements. Systems that exhibit stability are defined as *Normalized Systems* [20]. In such a system, the effort required for adding a function with given complexity does not depend on the overall system complexity. This is in contrast to the situation typically observed in software projects, where combinatorial effects or instabilities cause this relationship to become exponential (Figure 2).

Mannaert et al. also proposed that, in Normalized Systems, modular structures should strictly adhere to the following principles [21]:

1) Separation of Concerns: *An Action Entity can only contain a single task.*

   Each task must be able to evolve independently. If it is expected that two or more aspects of a program function (i.e., tasks) will evolve independently, they must be separated. It is proven that if one module contains more than one task, an update of one of the tasks requires updating all the others, too. Therefore, Normalized Systems shall be constructed of Action Entities (independent code modules) dedicated to one core activity.

   Most discussions regarding the separation of concerns remain vague about what a concern is actually. In contrast, the Normalized Systems theory provides an explicit definition by introducing the concept of so-called change drivers: A module should not contain parts that can evolve separately; rather, these parts should be placed in separate modules.

2) Data Version Transparency: *Data Entities that are received as input or produced as output by action entities need to exhibit Version Transparency.*

   Data Version Transparency requires that Data Entities (variables, records) can exist in multiple versions without affecting the actions that consume or produce them. In other words, an update of a Data Entity shall not affect the interface of an Action Entity, i.e., it must be possible to use different versions of this Data Entity in the same way when exchanging parameters or arguments with an Action Entity.

3) Action Version Transparency: *Action Entities that are called by other Action Entities need to exhibit Version Transparency.*

   Action Version Transparency implies that an Action Entity can have multiple versions, without affecting the way this Action Entity is invoked. In other words, introducing a new version of an Action Entity or task shall not require changes to any other Action Entities calling (or called by) the Action Entity containing the task.

4) Separation of States: *The calling of an Action Entity by another Action Entity needs to exhibit State Keeping.*

   This theorem focuses on the interaction of Action Entities with other Action Entities, more specifically, on the aggregation and mutual invocation of Action Entities in order to perform a function encompassing multiple tasks. State Keeping requires every Action Entity to be itself responsible for keeping track of its requests to other Action Entities. If results are not returned as expected, the calling action entity must not block indefinitely; rather, it shall handle the exceptional situation in an appropriate way.

V. Migration and rewrites

Already decades ago, Doug McIlroy called for software building blocks which can be safely regarded as black boxes [25]. Such blocks should not contain any undesired, hidden dependencies. A truly Normalized System fulfils this requirement. However, such a system must comply with the Normalized Systems theorems from the ground up, down to the smallest building blocks or modules. Since each Action Entity may only contain one task, this leads to very fine-grained structure with an enormous amount of very small modules.

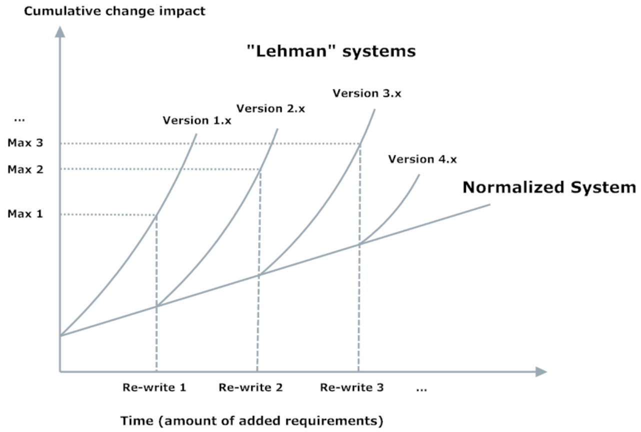


Figure 3. Reduction of cumulative effort by way of a rewrite

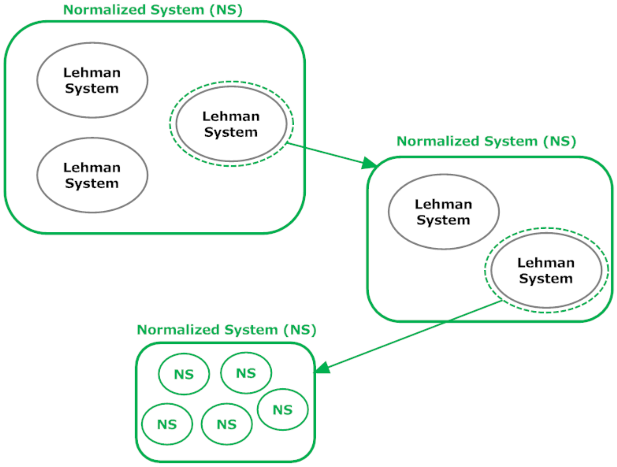


Figure 4. Migration from Lehman to Normalized subsystems

Restructuring a legacy (‘Lehman’) system into a Normalized System is, of course, a formidable task. A good start could be just separating the system into larger modules or subsystems with normalized interfaces. Each subsystem could then be considered to be a Normalized System, although it will likely still contain non-normalized subsystems (larger modules; see Figure 4). Such a subsystem module will not be a McIlroy-type safe black box due to its Lehman-type subsystems containing hidden dependencies. However, the combinatorial effects originating from its Lehman-type subsystems are stopped by the module interface, which complies with the Normalized Systems theorems.

In the case of automation components, web services (possibly making use of OPC UA) could be a very helpful mechanism to separate PLCs to transform them into subsystems with normalized interfaces. However, this is not enough. The internal modules of the subsystem, that is, all IEC 61131-3 code, must eventually be structured following the principles of Normalized Systems to make the subsystem a truly stable system.

When discussing the web as a design guide earlier in this paper, we observed that objects on the web, i.e., web sites, still can ‘crash’. The same is true for a PLC which uses code that does not follow the principles of Normalized Systems. Both the web site and the PLC are still complex subsystems that, eventually, need to be rewritten, probably by again splitting them into different subsystems.

In general, the first step in a migration scenario to let a non-normalized system evolve into a normalized one would therefore be to identify parts that can be readily isolated from the remaining parts. After adding a normalized interface, each of the isolated parts can be replaced by a normalized software re-write. Such a well-performed maintenance activity or ‘re-write’ will reduce the combinatorial effects within the system or subsystem (visible in Figure 3 as discontinuities along the y-axis). If full normalization is reached with the re-write, combinatorial effects will be removed entirely and permanently. Since the original part was already isolated via a normalized interface, the ‘version change’ brought about by the re-write will not cause combinatorial effects. Thus, once the parts of the overall integrated system have been isolated from each other (e.g., by way of web services), every Lehman-type subsystem can be transparently replaced by a Normalized one.

VI. Conclusion and Outlook

The timeline proves that the designers of the technologies behind the Internet applications we know today could not be aware of the Normalized Systems theory. However, the Internet as a system complies with the principles of this theory surprisingly well. Of course, the Normalized Systems theorems are not completely new, but have been available for a long time, albeit in the form of tacit knowledge. Those designers did a remarkable job following their intuition, and realised a world wide system that is stable even if the application objects within this system are not – thanks to loose coupling and excellent separation of concerns.

We are convinced that the Normalized Systems theory aids in achieving loosely coupled systems. It promotes the development of extremely fine-grained subsystems. As a first step on this way, isolating combinatorial effects in automation systems can be achieved by introducing web services based interfaces. These interfaces separate technologies, platforms, and vendor-dependent products; more generally spoken, they apply the Separation of Concerns principle to automation systems. OPC UA has a promising role in this regard thanks to the widespread adoption of OPC in industry.

Web-based or Web-enabled automation components can be regarded as a black box or isolated subsystem through an OPC UA interface [19]. If we manage to find concepts to restructure IEC 61131-3 code into very many small components, the IEC 61131-3 information model OPC UA companion standard could be applied to make this structure transparent. We do not see a reason why the amount of such subsystems, interconnected via web services, would be limited, except for limited hardware resources. Having a large region system, comparable with the Internet, that interconnects automation control subsystems, might become very valuable in the future smart grid. Indeed, the Internet, which is mainly used for interconnecting information sources, might be extended with OPC UA based capabilities for interconnecting production control resources.

Certainly, it will be essential to further investigate which mechanisms the web and web services are built on that are responsible for the desirable system properties mentioned and translate them into the industrial automation domain.

REFERENCES

[1] T. Sauter and M. Lobashov, “How to Access Factory Floor Information Using Internet Technologies and Gateways,” IEEE Trans. Ind. Inform., vol. 7, no. 4, Nov. 2011, pp. 699–712.

[2] T. Lewis, “Netstorms: the crash of ’96,” Computer, IEEE Computer Society , vol. 29, no. 11, Nov. 1996, pp. 12–14.

[3] R. Cohen, K. Erez, D. ben-Avraham, and S. Havlin (2000), “Resilience of the Internet to random breakdowns,” Physical Review Letters, vol. 85, no. 21, Nov. 2000, pp. 4626–4628.

[4] R. Baldoni, S. Bonamoneta, and C. Marchetti, “Implementing Highly-Available WWW Servers based on Passive Object Replication,” 2^{nd} IEEE International Symposium on Object-Oriented Real-Time Distributed Computing (ISORC ’99), Saint-Malo, 1999, pp. 259–262.

[5] D. Kelly, “A study of design characteristics in evolving software using stability as a criterion,” IEEE Transactions on Software Engineering, vol. 32, no. 5, pp. 315–329, May 2006.

[6] L.B. Jackson, “Digital Filters and Signal Processing,” 2^{nd} ed., Kluwer Academic Publishers, Norwell, MA, 1988.

[7] A. V. Oppenheim, A. S. Willsky, and S. H. Nawab, “Signals & Systems,” 2^{nd} ed., Prentice-Hall, Upper Saddle River, NJ, 1996.

[8] H. P. Breivold, I. Crnkovic, R. Land, and M. Larsson, “Analyzing Software Evolvability of an Industrial Automation Control System: A Case Study,” The Third International Conference on Software Engineering Advances, Oct. 2008 (ICSEA ’08), pp. 205–213, 2008.

[9] M. M. Lehman, “Programs, life cycles, and laws of software evolution,” Proceedings of the IEEE, vol. 68, pp. 1060–1076, 1980.

[10] International Electrotechnical Commission, “IEC 61131-3, Programmable controllers – Part 3: Programming languages,” 2003.

[11] D. van der Linden, H. Mannaert, and P. De Bruyn, “Towards the Explicitation of Hidden Dependencies in the Module Interface,” ICONS 2012, 7^{th} International Conference on Systems, accepted for publication, 2012.

[12] D. van der Linden, H. Mannaert, and J. de Laet, “Towards evolvable Control Modules in an industrial production process”, ICIW 2011, 6^{th} International Conference on Internet and Web Applications and Services, pp. 112–117, 2011.

[13] F. Jammes, H. Smit, “Service-Oriented Paradigms in Industrial Automation,” IEEE Transactions on Industrial Informatics, vol. 1, no. 1, 2005.

[14] P. Spiess et al., “SOA-based integration of the Internet of Things in Enterprise services,” IEEE International Conference on Web Services, Los Angeles, USA, July 6–10, 2009.

[15] T. Hannelius, M. Salmenpera, and S. Kuikka, “Roadmap to adopting OPC UA,” 6^{th} IEEE International Conference on Industrial Informatics, pp. 756–761, 2008.

[16] A. P. Kalogeras, J. V. Gialelis, C. E. Alexakos, M. J. Georgoudakis, S. A. Koubias,“Vertical Integration of Enterprise Industrial Systems Utilizing Web Services,” IEEE Transactions on Industrial Informatics, vol. 2, no. 2, 2006.

[17] J. Lange, F. Iwanitz, and T. J. Burke, “OPC: von Data Access bis Unified Architecture”, VDE-Verlag, 2010.

[18] W. Mahnke, S. Leitner, and M. Damm, “OPC Unified Architecture”, Springer, 2009.

[19] D. van der Linden, H. Mannaert, W. Kastner, V. Vanderputten, H. Peremans, and J. Verelst,“An OPC UA Interface for an Evolvable ISA88 Control Module,” ETFA 2011, IEEE Conference on Emerging Technologies and Factory Automation, 2011.

[20] H. Mannaert and J. Verelst, ”Normalized Systems Re-creating Information Technology Based on Laws for Software Evolvability,” Koppa, 2009.

[21] H. Mannaert, J. Verelst, and K. Ven, “The transformation of requirements into software primitives: Studying evolvability based on systems theoretic stability,” Science of Computer Programming, 2010.

[22] OPC Foundation, “OPC Unified Architecture,” www.opcfoundation.org.

[23] International Electrotechnical Commission, IEC 62541-1, “OPC unified architecture – Part 1: Overview and Concepts,” 2010.

[24] D. van der Linden and H. Mannaert, “In Search of Rules for Evolvable and Stateful run-time Deployment of Controllers in Industrial Automation Systems,” ICONS 2012, 7^{th} International Conference on Systems, Reunion, pp. 67-72, 2012.

[25] M. D. McIlroy, “Mass produced software components,” NATO Conference on Software Engineering, Scientific Affairs Division, 1968.

Preliminary Ideas on Concept to Query Closeness Metrics

Alejandra Segura N.
Depto. Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
e-mail: asegura@ubiobio.cl

Christian Vidal-Castro.
Depto. Sistemas de Información
Universidad del Bío-Bío
Concepción, Chile
e-mail: cvidal@ubiobio.cl

Claudia Martinez A.
Depto. Ingeniería Informática
Universidad Católica de la Santísima Concepción
Concepción, Chile
e-mail: cmartinez@ucsc.cl

Salvador Sánchez-Alonso.
Depto. Ciencias de la Computación
Universidad de Alcalá
Madrid, España
e-mail: salvador.sanchez@uah.cl

***Abstract*—The usefulness of knowledge models for information retrieval tasks such as digital resource tagging, query expansion, and recommending, among others, requires that the query concept be present in the model, i.e., exists matching. If the query concept is absent in the ontology, exists matching problem. In this case, it can be identified in the model other sematic and syntactically closeness concepts to the query concept. Once identified the closest concept is possible to extract relevant knowledge from the ontology. The goal of this work is to propose a solution to the problem mentioned, by identifying those variables that can affect the closeness between a query and concepts in a domain ontology. Using these variables as a starting point, we propose 6 indexes for measuring the degree of closeness. We present the results of implementing a search-selection algorithm using indexes based on exact words, contained words, coincidences in descriptive fields, new words and approximate depth. These indices are validated via a case study, and, from these results, we recommend adjustments needed for building a global concept closeness index in future works.**

Keywords-component; Ontology; Semantic Web; Web Information Retrieval.

I. INTRODUCTION

Information retrieval (IR) involves several processes, among which we can distinguish indexing, query, search and relevance assessment [1].

Knowledge models, mainly thesauri, terminologies and ontologies, provide external knowledge that can semantically enrich, either directly or indirectly, several tasks related to information retrieval, such as digital resource tagging, indexing, querying and recommending. For example, indexing can either build an index by extracting information for resource tagging or it can use the knowledge model itself as an indexing system [2, 3]. Knowledge models are used in the formulation and refinement processes to navigate among the modeled concepts, and also in query expansion to disambiguate or further specify the initial user query by adding new information to the query [4, 5]. In relevance assessment, knowledge models have been used to rank results according to relevance, in what is called “score of results” [6].

Knowledge models can be used either manually, i.e. the user defines the query through model navigation, or automatically, through the use of algorithms that extract relevant information. Despite the fact that manual extraction can yield more precise results, its application is limited due to the large size and structure of many models and because, in most cases, users must have previous knowledge of the model [7]. On the other hand, automatic knowledge extraction allows using large knowledge models and makes their structure transparent to the users. However, their use is generally restricted to those cases where the query is exactly represented in the model.

According to [8], an ontology is “an explicit specification of a conceptualization”. Ontologies involve two parts: syntax and semantics. The first considers symbols and the set of rules for combining them, and the second refers to the meaning of expressions. Ontologies rigorously specify a conceptual framework in a domain, with the goal of facilitating communications, interaction, exchange and information sharing between different computational systems.

Knowledge representation, therefore, requires domain knowledge, representation languages and mechanisms for inferring new knowledge. As indicated in [9], ontologies are the tool of choice for formal knowledge representation oriented to computer-assisted semantic analysis.

A problem associated to the use of knowledge models as a basis for automatic knowledge extraction occurs when an exact match to the query cannot be found. Then, the knowledge model can be examined looking for concepts that are closely related to the query. Accessing the concept closest to the query in turn makes it possible to access other semantically related concepts. That is, new knowledge can be extracted and then utilized in any other information retrieval processes, such as digital resource tagging, indexing, recommending or query expansion.

This article describes an algorithm that, given a query, extracts those concepts that are closest to the query from a given ontology. It also presents an analysis of the proposed algorithm's initial assessment.

The remainder of the article is organized as follows. Section II formalizes the problem considering the non-exact correspondence between the query and the knowledge modeled in the ontology. Section III analyzes previous works related to syntactic and semantic similarity metrics, and also to ontology-based query expansion algorithms. Section IV describes the research methodology, while Section V describes the evaluation process, which includes validation by experts, and the design and application of a questionnaire. Section VI analyzes and discusses the results. Finally, Section VII presents our conclusions and future research directions.

II. The query matching problem

In this section, we define the basic elements of the matching problem, which are the query concept and the domain ontology.

Query Concept: The user's query is the query concept (QC), which is formed by a set of words w, that is, QC = $\{w_1, \ldots, w_n\}$. Let QC' be the same query concept after linguistic processing, that is, after removing morphological variations (stemming) and ignoring stopwords. Common, frequently-used words generally do not provide information and thus are considered stopwords. The set of stopwords includes prepositions, articles, adverbs, conjunctions, possessive and demonstrative pronouns, and some verbs and nouns. Stopword lists are generally language dependent, but some domain-dependent stop-word lists have also been built [10]. Stemming is the process by which morphological variations of the terms are extracted, e.g., conjugations as well as prefix and suffix derivational morphemes. A derivational morpheme is appended or prepended to a lexical base to form a new derived word. Eliminating these morphemes leaves only the root. Therefore, the lemma represents the variations of the derived terms [11].

Domain ontology: in this work, based on [12], we define a domain ontology as a triplet O={C, R, I}, where C is the set of classes, R is the set of relationships between classes and instances, and I is the set of class instances. Any concept modeled in the ontology is represented either in the classes or in the instances. Every ontology modeled concept (OC) is a set of words such that OC= $\{v_1, \ldots, v_y\}$. The ontologies have relationships related to concept taxonomies such as *is-a* or *part-of,* even though they can also include domain-specific relationships to take into account the modeling requirements of the knowledge domain.

Considering the above definitions, the matching problem between a query concept and the concepts modeled in a domain ontology exists when:

$$\nexists OC \in O : QC' = OC$$

Most of the work in information retrieval that makes use of knowledge models assumes that there is a matching between the query concept and at least one concept in the ontology. Although the query concept is absent from the ontology, can be identified in the model other closeness concepts to the query concept (QC). The degree of closeness to the query concept might be determined according to syntactic and semantic variables. It should be noted that there is little information about the query concept context to determine the closeness between the query concepts and the concepts in the ontology. Specifically, we only know the concept (and set of words) and the domain of knowledge where it is immersed.

Our proposal presents an algorithm aimed at extracting those concepts in the ontology which are closest to a query concept for which no exact match exists.

III. Related work

The problem of matching a query to a domain ontology has been studied in relatively few ontology-based query expansion algorithms, most of which perform the query expansion only if the query concept exists exactly in the model, that is, there is a concept which contains the same words keeping the same order [3, 4, 13-17]. Moreover, our study of related work also reviews relevant syntactic and semantic similarity measures, as they indirectly affect the problem at hand. The similarity has been managed both syntactically and semantically. The syntactic aspect is based on the comparison of two strings and the semantic aspect through the comparison between two concepts present in the model. In the latter case exists two approaches: one based on the structure and another based on the information content.

The edit distance measure (e) proposed by [18] is used to determine the degree of syntactic similarity between 2 strings A and B. It is defined as the number of removal, replacement or append operations needed to convert string A into string B.

These semantic similarity measures consider that both concepts are represented in the model. Other structure-based measures use as a basis the number of nodes separating both concepts. Proposed measures utilize variables such as the depth of the lowest common ancestor (LCA), the local density of the sub-tree containing both concepts, the distance between the concepts and the types of relationships among them. For example, the measure proposed by [19] is calculated as the shortest route between the concepts. The measure proposed by [20] is calculated as a function of the depth of the LCA and the number of links between the concept and said ancestor. The similarity measure proposed by [21] is a function of the concepts' depth, the depth of the LCA and the shortest distance between the concepts. The same authors propose another similarity measure that also includes the local specificity [22]. Li, et al. [23] propose a similarity measure that takes into account the shortest route between the concepts, the depth of the LCA and empirical information.

Information-based semantic similarity measures consider the information content of the model's derived nodes and corpus statistics, such as the concept's frequency in the corpus and the corresponding inverse frequency. The more information two concepts share, the higher their similarity. Some similarity measures in this category are the ones proposed by Resnik [24], which take the information content of the LCA into account. Jiang and Conrath [25] and Lin

[26] suggest improvements to Resnik's measure which also consider the information content of each concept.

The above mentioned semantic and syntactic similarity measures are not directly applicable to the correspondence problem, as they can be used only if both concepts are present in the model. In our case, however, the query concept is absent from the model. Nevertheless, these works are relevant to formulating our proposed solution.

Previous work proposes a query expansion algorithm based on domain ontologies [5]. The same work also defines an algorithm for finding the concept closest to a query concept not present in the ontology. The closest concept is defined as the concept that contains the largest number of words in common with the query concept, and that contains the smallest number of words that do not belong to the query concept.

IV. METHODOLOGY

Figure 1 shows the framework that describes how the matching problem is addressed. In any IR process, when the query concept is absent in the model, it is processed linguistically. Then, the concepts that share words with the query concept are extracted from the ontology. These concepts are also processed linguistically to calculate the indexes of closeness. Finally, based on the indexes, the global concept closeness index is estimated.

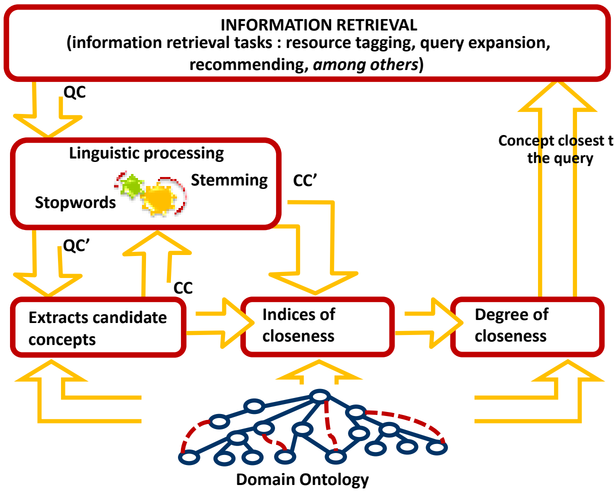

INFORMATION RETRIEVAL
(information retrieval tasks : resource tagging, query expansion, recommending, among others)
QC
Concept closest to the query
Linguistic processing
Stemming
Stopwords
CC'
QC'
CC
Extracts candidate concepts
Indices of closeness
Degree of closeness
Domain Ontology

Figure 1. Matching problem framework.

We define the candidate concepts (CC) for a given query concept (QC) as all those concepts present in the ontology that share words with the linguistically pre-processed query QC'. Linguistic processing includes stopword elimination and plural extraction: in other words, a full stemming process is not performed.

Let QC_i be a query concept composed of a set of words w, and QC'_i is the linguistically preprocessed query. Also, let OC be any concept modeled in a domain ontology, which in turn is defined as a set of words v_{nt} such that $OC_1 = \{v_{11}, \ldots, v_{1p}\}, \ldots, OC_n = \{v_{n1}, \ldots, v_{nt}\}$. Then, if QC_i is not present in the ontology, we can say that OC_n is a candidate concept for QC'_1 if and only if

$$\exists v_{nt} \in OC_n \wedge \exists w'_{ip} \in QC'_i \,/\, v_{nt} = w'_{ip}$$
$$\wedge\ v_{nt} \neq stopword \wedge$$
$$\wedge\ v_{nt} = stemm(v_{nt})$$

The closeness from the query concept QC' and the candidate concepts CC is a function of the following

1.-The number of words that match the query concept words. Two types of coincidences, exact and contained, are considered.

> 1.1.- A contained coincidence occurs when the query concept word is contained within a candidate concept's word.
>
> 1.2.- An exact coincidence occurs when the query concept word is syntactically identical to a word in the candidate concept. The greater the number of coincident words, the closer the query concept and the candidate concept.

2.- Word positions. In addition to the coincidences among the query concept's words and the candidate concepts' words, the coincident word's position is also considered. Analysis of this parameter can vary according to each language's grammatical rules. A candidate concept's closeness to the query concept increases if word positions also coincide.

3.- Number of new or mismatching words. This criterion counts the number of CC words that do not coincide either exactly or approximately with any query concept word. Stop-words are ignored. The fewer the new or mismatched words, the higher the candidate concept's closeness to the query concept.

4.- Concept depth. The depth is defined as the longest path from the candidate concept to the model's root class, considering hierarchical is-a relationships. In a domain ontology, the deeper the concept the more specific it is.

5.- Parent relevance. This item considers the parents of the candidate concepts and quantifies the number of its descendants that are also candidate concepts. Closeness increases if a candidate concept belongs to a sub-tree with a greater candidate concept density.

6.- Descriptive fields representation. This item considers the occurrence of the query concept in any descriptive field associated to the candidate concept, such as the <definition> or <description> fields. Concept closeness increases if the candidate concept's descriptive fields contain the query concept.

The 7 variables just mentioned are considered relevant for determining the closeness between a query concept and those concepts modeled in an ontology. Next, we show the 6 indices to be calculated by the algorithm for each candidate concept. Each index takes values between 0 and 1.

Normalized exact word index

$$ind_{coin_{ex}} = \left(\frac{c_{word_{ex}}}{t_{word_{qc}}} \right) \qquad (1)$$

where:

c_word_ex : Number of query words that are also present, exactly, in the candidate concept (variable 1.1).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

ISBN: 978-1-61208-200-4

Normalized contained word index

$$ind_{coin_{co}} = \left(\frac{c_{word_{co}}}{t_{word_{qc}}} \right) \quad \textbf{(2)}$$

where:

c_word_co : Number of query words that are contained in a word present in the candidate concept (variable 1.2).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

Normalized new word index

$$ind_{nue} = \left[1 - \frac{t_{word_{cc}} - (c_{word_{co}} + c_{word_{ex}})}{t_{word_{cc}}} \right] \quad \textbf{(3)}$$

where:

c_word_ex : Number of words in the query concept that are also present in the candidate concept (variable 1.1).
c_word_cc : Number of words in the query concept that are contained in a word present in the candidate concept (variable 1.2).
t_word_cc : Total number of words in the candidate concept CC, ignoring stopwords
t word cc –(c_word_cc + c_word_ex): Number of new or mismatching words (variable 3).

Descriptive fields coincidence index

$$ind_{coin_{des}} = \left(\frac{c_{word_{des}}}{t_{word_{qc}}} \right) \quad \textbf{(4)}$$

where:

c_word_des : Number of words in the query concept that are an exact or partial match to words in the candidate concept's descriptive fields (variable 6).
t_word_qc : Total number of words in the query concept QC, ignoring stopwords.

Normalized aproximate depth index.

Given the computational complexity inherent to the problem of exactly calculating a concept's maximum depth in a formal domain ontology, this index is defined as an approximation to the candidate concept's maximum depth (variable 4). A formal domain ontology usually includes a large number of modeled concepts, each of which can have several parent nodes, therefore many routes to the root node can exist.

To calculate the semantic distance each line of inheritance has value 1. Therefore we assume that any inheritance relationship with a class that belongs to another ontology is assigned half this value (t_subclassof_o=0.5). This avoids calculating the depth in external ontologies.

$$ind_{prof_{app}} = \left(\frac{ind_{prof}}{\max\limits_{cc \ of \ qc} (ind_{prof})} \right)$$

$$ind_{prof} = \left(\frac{t_{ant cc}}{p_{subclassof_{o}}} \right) \quad \textbf{(5)}$$

$$p_{subclassof_{o}} = \left(\frac{t_{subclassof_{o}}}{t_{class_{o}}} \right)$$

where:

p_subclassof_o : parent relationship average for all concepts in the ontology.
t_subclassof_o : total number of subclass relationships in the ontology. Subclass relationships have weight 1, while references to classes in other ontologies or sub-ontologies have weight 0.5.
t_class_o : total number of classes modeled in the ontology that have at least one parent (except for root nodes).
t_ant_cc : total number of parents of a candidate concept.
max(cc of qc): maximum approximate depth of all candidate concepts for a given query.

Candidate density in sibling index (variable 5).

$$ind_{den} = \max_{parent \ CC} \left(\frac{c_{sibling_{CC}}}{t_{sibling_{CC}}} \right) \quad \textbf{(6)}$$

where:

max parent cc : maximum value among all the candidate concept's parents.
parent_cc : number of concept candidate's parents.
c_sibling_cc : number of candidate concept's siblings that are also candidate concepts.
t_sibling_cc : total number of candidate concept's siblings.

V. Evaluation

We wanted to evaluate the algorithm by examining the concepts it retrieves and determining their closeness to the query concept. The algorithm is evaluated using the Subcellular Anatomy for the Nervous System (SAO) ontology, which is available in the OWL language. This ontology provides a method for describing sub-, supra- and macro-cellular structures. SAO "describes the parts of neurons and glia and how these parts come together to define supracellular structures such as synapses and neuropil", and was developed by the Open Biological and Biomedical Ontology Foundry (http://www.obofoundry.org/crit.shtml) with the stated aim of providing updated domain ontologies in several knowledge areas for the scientific community [27].

For evaluation purposes, we utilize test queries extracted from the syllabi of four central nervous system Anatomy courses. Query concepts are extracted from the contents list for each course. Details can be found in Table I.

TABLE I. SYLLABI USED FOR ALGORITHM EVALUATION

Course details
Learning and Memory: Activity-Controlled Gene Expression in the Nervous System. Fall 2009 . http://ocw.mit.edu/courses/biology/7-340-learning-and-memory-activity-controlled-gene-expression-in-the-nervous-system-fall-2009/Syllabus/
Psychology 202 Biopsychology. Fall 2009. http://courses.washington.edu/psy222/Syllabi/Psy%20202%20Fall%2009%20syl.pdf
Neurophysiology 1012 and 2012. Spring 2009. http://www.neuroscience.pitt.edu
Neuro 405- Neurophysiology .Fall 2010. http://webpub.allegheny.edu/employee/l/lfrench/Neurophys%20syllabus%20F06.htm

73 initial query concepts were identified. For each initial query concept, the algorithm generated a list of candidate concepts sorted by closeness, according to the scores of the 6 indices mentioned in Section 4.

Faculty from the Universidad Católica de la Santísima Concepción, with professional experience in medicine, specifically in anatomy and cellular biology, participated as experts in this study. At first, these experts were consulted to filter the initial concepts and to select those that were coherent with their research lines. Three experts agreed in the selection of 7 initial query concepts. Table II details algorithm results for these 7 queries.

TABLE II. LIST OF INITIAL QUERY CONCEPTS USED IN THE FIRST PHASE OF THE EVALUATION WITH THE RESULTS OF THE ALGORITHM

	Indices												
		coinex		coinco		nue		coindes		prof_app		den	
Queries	**t**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**	**A**	**B**
Activation of the NMDA receptor	30	0	0.67	0	0.67	0	1	0.53	1	0	0.67	0.04	1
AMPA receptor endocytosis	31	0	0.67	0	0.67	0	1	0.41	1	0	0.67	0.04	1
APs- Ca channels	24	0	0.33	0	0.33	0	0.5	0.47	1	0	1.33	0.11	1
Brain	17	0	0	0	1	0	0.33	0.13	1	1	2	0.03	1
cerebro spinal fluid	11	0	0.33	0	0.33	0	0.33	0.13	1	0.33	1	0	1
Electrical principles of neuronal function	49	0	0.25	0	0.25	0	0.5	0.27	1	0	0.5	0	1
Neurons	75	0	1	0	1	0	1	0.12	1	0	2	0	1

t: total number of candidate concepts
coin_ex: normalized exact word index.
coin_co: normalized contained word index.
coin_des: descriptive fields coincidence index.
nue: normalized new word index.
prof_app: normalized approximate depth index.
den:candidate density in sibling index.
A: minimun.
B: maximum.

Each expert evaluated the first 10 candidate concepts, randomly sorted, for each of the 7 initial queries through a questionnaire named “Concept closeness evaluation in a domain ontology”. This instrument was designed to gather expert opinions regarding:

- the conceptual closeness of the initial query concept to the candidate concept,
- the closeness rank of the 10 candidate concepts, a number from 1 to 10, where 10 denotes greatest closeness.

In the first phase of the evaluation, a single expert was chosen so as to do an exploratory case study. This was done to find out “the relationship between the closeness ranking determined by the expert and the indices computed for each candidate concept by our algorithm”.

The expert evaluated the closeness of 10 candidates for 7 initial queries. Of these 70 measurements, 64 candidate concepts (91%) were considered close and only 6 (9%) were found to be not close or unrelated to the initial query. The expert indicated that the strategy he applied was, after determining closeness, to perform a top-down revision based upon his experience with the candidate concepts in terms of their composition relationships.

VI. RESULTS AND DISCUSSION

These results were analyzed using Pearson correlation analysis [28]. This Pearson analysis was performed to find correlations between the closeness rank specified by the expert (on a scale of 1 to 10, 10 denoting greatest closeness) and each of the indices proposed in Section 4. Analysis results are shown in Table III. Five of the correlation indices were found to be positive relationships, that is, a higher expert ranking yields a larger estimated index. Pearson coefficients concentrated in the (0.46, 0.07) range.

The best correlation values for the closeness rank were obtained for the new word index (0.46), the exact word index (0.40) and the contained word index (0.40). On the other hand, the least relevant correlation was obtained for the descriptive fields coincidence index (0.07). This low correlation can be explained by considering the poor structure and flexibility allowed when filling these descriptive fields. Additionally, it must be considered that the goal of these fields is mainly to provide information to other users.

TABLE III. THE CLOSENESS RANKING DETERMINED BY THE EXPERT AND THE RANKING GIVEN BY THE ALGORITHMS AND THE CALCULATED INDICES.

ordex-coinex	ordex-app	ordex-nue	ordex-def	ordex-prof	ordex-dens
0.40	0.40	0.46	0.07	-0.02	0.19

ordex: closeness ranking determined by the expert
coin_ex: normalized exact word index.
coin_co: normalized contained word index.
coin_des: descriptive fields coincidence index.
nue: normalized new word index.
prof_app: normalized approximate depth index.
den: candidate index in sibling index.

The only index that showed a negative low correlation was the approximate depth index, with a value of -0.02. Beforehand, we expected a higher positive correlation, under the premise that candidate concepts that are deeper in the ontology are more specific, which would in turn yield a higher closeness rank (closer to 10) with respect to the query concept. However, the data shows that the deeper the depth index the lower the closeness rank is, i.e. the candidate concept has a ranking closer to 1. This can be explained by noting that the query concepts (content lists of a course) and the concepts in the ontology have differences in the granularity/specialization. The query concept was assumed to be very specific, so then a deep candidate concept would be very close. However, not all query concepts are specific. Then, if the query concept is of a general nature, its closest candidate concepts will also be of a general nature. Therefore, the relevance of the depth index depends on whether the query concept is of a general or a specific nature. Unfortunately, as the query concept is not present in the ontology, we do not have information about its depth.

In general terms, our results show that the indices proposed in this work are useful as a measure of the closeness between a query concept and concepts modeled in an ontology. As such, they can be used as a starting point for the development of a global closeness index that can be used to rank those concepts that are closest to the query concept.

VII. Conclusions and Outlook

The usefulness of knowledge models for information retrieval tasks such as digital resource tagging, query expansion, and recommending, among others, requires that the query concept be present in the model. This work addresses the matching problem that occurs when the query concept is not present exactly in the model. We postulate that it is possible to find concepts that are syntactically and/or semantically close to the query concept, even if the query is not represented in the ontology, and that the closeness between the query concept and a candidate concept can be determined as a function of 7 variables. Based on these 7 variables, we define 6 normalized indices for estimating concept closeness, which are the exact word index, the descriptive fields coincidence index, the contained word index, the new word index, the approximate depth index, and the candidate index in siblings index. After a first evaluation phase, we conclude that 5 of the 6 indices are positively correlated with the closeness rank perceived by domain experts. Moreover, one of the proposed indices warrants further research as its incidence on closeness rank depends on the generality or specificity of the query concept. This, in turn, leads us to envision a mechanism that allows knowing a priori a query concept's depth so as to be able to calibrate the candidate concepts' closeness rank.

As future work, we must determine the degree of incidence define in the closeness rank estimation. In order to do this, we will perform a new evaluation with a larger number of experts, and also we will consider changing the knowledge domain area so as to generalize the results obtained to date.

Acknowledgments

This work was partially supported by DIUBB 115215 3/RS Chile; Bío-Bío University, Chile. Some research in ontologies is also funded by Organic.Lingua CIP-ICT-PSP.2010.6.2 project, reference number 270999.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*: Addison-Wesley-Longman, 1999.

[2] C. D. Nguyen, K. J. Gardiner, and K. J. Cios, "Protein annotation from protein interaction networks and Gene Ontology (In press)," *Journal of Biomedical Informatics,* p. 6, 2011.

[3] G. Zou, B. Zhang, Y. Gan, and J. Zhang, "An Ontology-Based Methodology for Semantic Expansion Search," in *FSKD '08: Proceedings of the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery*, 2008, pp. 453-457.

[4] Y.-F. Huang and C.-H. Hsu, "PubMed smarter: Query expansion with implicit words based on gene ontology," *Knowledge-Based Systems.,* vol. 21, pp. 927-933, 2008.

[5] A. Segura N., S. Sánchez, E. García-Barriocanal, and M. Prieto, "An empirical analysis of ontology-based query expansion for learning resource searches using MERLOT and the Gene ontology," *Knowledge-Based Systems,* vol. 24, p. 15, 2011.

[6] F. Farfan, V. Hristidis, A. Ranganathan, and M. Weiner, "XOntoRank: Ontology-Aware Search of Electronic Medical Records," in *Proceedings of the 2009 IEEE International Conference on Data Engineering*: IEEE Computer Society, 2009.

[7] J. Bhogal, A. Macfarlane, and P. Smith, "A review of ontology based query expansion," *Information Processing and Management: an International Journal,* vol. 43, pp. 866-886, 2007.

[8] T. R. Gruber, "A translation approach to portable ontology specifications," *Knowledge Acquisition,* vol. 5, pp. 199-220, 1993.

[9] M.-Á. Sicilia, E. García-Barriocanal, C. Pages, J. Martinez, and J. Gutierrez, "Complete metadata records in learning object repositories some evidence and requirements," *International Journal of Learning Technology,* vol. 1, pp. 411-424, 2005.

[10] Á. F. Zazo, C. G. Figuerola, J. L. Alonso Berrocal, and R. Emilio, "Reformulation of queries using similarity thesauri," *Information Processing and Management: an International Journal,* vol. 41, pp. 1163-1173, 2005.

[11] M. F. Porter and K. W. P. Sparck-Jones, *An algorithm for suffix stripping*: Morgan Kaufmann Publishers Inc., 1997.

[12] K. Todorov, P. Geibel, and K.-U. Khnberger, "Mining concept similarities for heterogeneous ontologies," in *Proceedings of the 10th industrial conference on Advances in data mining: applications and theoretical aspects* Berlin, Germany: Springer-Verlag, 2010.

[13] A. Abdelali, J. Cowie, and H. S. Soliman, "Improving query precision using semantic expansion," *Information Processing and Management: an International Journal,* vol. 43, pp. 705-716, 2007.

[14] M. C. Díaz-Galiano, M. T. Martín-Valdivia, and L. A. Ureña-López, "Query expansion with a medical ontology to improve a multimodal information retrieval system," *Comput. Biol. Med.,* vol. 39, pp. 396-403, 2009.

[15] M.-C. Lee, K. H. Tsai, and T. I. Wang, "A practical ontology query expansion algorithm for semantic-aware learning objects retrieval," *Computers &amp; Education,* vol. 50, pp. 1240-1257, 2008.

[16] L. Ma, L. Chen, Y. Gao, and Y. Yang, "Ontology Based Query Expansion in Vertical Search Engine," in *FSKD '09: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, 2009, pp. 285-289.

[17] J. Tuominen, T. Kauppinen, K. Viljanen, and E. Hyonen, "Ontology-Based Query Expansion Widget for Information Retrieval," in *Proceedings of the 5th Workshop on Scripting and Development for the Semantic Web (SFSW 2009), 6th European Semantic Web Conference (ESWC 2009)*, 2009.

[18] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady,* vol. 10, p. 3, 1996.

[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," *IEEE Transactions on Systems, Man and Cybernetics,* vol. v, pp. 17-30, 1989.

[20] Z. Wu and M. Palmer, "Verbs semantics and lexical selection," in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics* Las Cruces, New Mexico: Association for Computational Linguistics, 1994.

[21] H. Al-mubaid and H. A. Nguyen, "A Cluster-Based Approach for Semantic Similarity in the Biomedical Domain," in *The 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, San Francisco, CA, 2006.

[22] H. A. Nguyen and H. Al-Mubaid, "New ontology-based semantic similarity measure for the biomedical domain," in *IEEE International Conference on Granular Computing,* 2006, pp. 623-628.

[23] Y. Li, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources," *Knowledge and Data Engineering, IEEE Transactions on,* vol. 15, pp. 871-882, 2003.

[24] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume*

1 Montreal, Quebec, Canada: Morgan Kaufmann Publishers Inc., 1995.

[25] J. J. Jiang and D. W. Conrath, "Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy," in *International Conference Research on Computational Linguistics (ROCLING X)*, Taiwan., 1997, p. 15.

[26] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*: Morgan Kaufmann Publishers Inc., 1998.

[27] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotech,* vol. 25, pp. 1251-1255, 2007.

[28] J. Rodgers and A. Nicewander, "Thirteen Ways to Look at the Correlation Coefficient," *The American Statistician,* vol. 42, pp. 59-66, 1988.

Multi-dimensional Ontology Model to Support Context-aware Systems

A Mobile Application to Search and Invoke Semantic Web Services

José Rodríguez
Computing Department
CINVESTAV-IPN
DF, Mexico
rodriguez@cs.cinvestav.mx

Maricela Bravo
Systems Department
UAM-Azcapotzalco
DF, Mexico
mcbc@correo.azc.uam.mx

Rafael Guzmán
Computing Department
CINVESTAV-IPN
DF, Mexico
rguzman@computacion.cs.cinvestav.mx

Abstract— **Computing has evolved in such a way that has left behind the limitation of being static and local, promoting the emergence of the new era of computing called Ubiquitous Computing. To maximize the use of this new computing platform, it is necessary to develop new and improved structures for knowledge and information representing, in order support the implementation of new intelligent searcher and recommendation systems. Thus, recommendations and search results will be fully based on contextual information and user profiles. This paper describes an architecture based on a multi-dimensional ontology model to represent mobile user contexts, Web services and application domains. A mobile application is also presented to show the benefits of this approach in a mobile computing environment.**

Keywords-mobile computing; ontologies; semantic Web services; context-aware systems.

I. Introduction

Computing has evolved in such a way that has left behind the limitation of being static and local, that is, not being able to easily carry computing resources from one place to another due to its large size and energy demand. Currently this limitation has been solved with the reduced size, weight and power consumption of computing devices, resulting in the emergence of mobile computers, enabling communications between mobile devices and people anywhere in the world [1]. The mobile computing has become a reality, thanks to the evolution of wireless communication technologies and the Internet. Mobile computing is rapidly gaining importance because there is an incremental daily demand for information access from anywhere and at any time with multiple purposes. Nowadays, information is distributed around the world in multiple servers and databases, therefore to discover and gather this information accurately in mobile devices represents a big computational challenge.

Web services (WS) [2] represent an important technological trend for distributing software resources around the world. WS are software component interfaces based on a set of XML-based standards, languages and protocols and are executed through the exchange of XML messages, allowing different applications to communicate across multiple platforms. WS enable reutilization of legacy software and integration of more complex systems which can be of major interest in mobile environments.

Currently, there is a trend towards digital convergence, so it is possible to access information (voice, video or data) with any mobile devices via the Internet, which allows use and consumption of Web services. Emergent technologies for mobile phones are expected to offer full support for the Web, all telephone devices will have Internet access (generation 4G) [3].

These advances in telecommunications in turn promote the increased use of mobile devices in everyday life, so that computing power is distributed throughout the environment, becoming an invisible and integral part of our lives. This situation gives rise to the new era of computing called Ubiquitous Computing, which is considered the third generation of computing [4]. The main goal of Ubiquitous Computing is to integrate computers and devices in a physical environment of users trying to fully exploit the services offered by this new computing paradigm.

To maximize the use of this new ubiquitous computing platform, it is necessary to develop new and improved structures for representing information and knowledge to support the implementation of new intelligent search and recommendation systems. Thus, recommendations and search results will be fully based on contextual information and user profiles. Context is defined as "any information that can be used to characterize the situation of a person, place, or object that is considered relevant to the interaction between a user and application" [5]. In this way, systems that are able to extract, interpret and use information from the environment are known as context-aware systems.

It is therefore necessary to create a solution capable of taking into account the distribution of information, user mobility and consequently the context information of the user, enabling access to Web services offered on the network.

The rest of the paper is organized as follows: Section II presents a motivation example; in Section III, related work is presented; in Section III, the multi-dimensional ontology model is described; in Section IV, a Web service semantic registry is presented; Section V, describes a context-aware mobile application; in Section VI, experiments and results are presented; and finally, in Section VII, conclusions are described.

ISBN: 978-1-61208-200-4

II. MOTIVATING EXAMPLE

To provide an example of the approach presented in this article, consider a mobile application that allows any user to connect to a public Semantic Web Service Registry (SWSR) to search for web services and invoke them as needed. During the registration of the mobile user, the SWSR obtains the mobile user data (full name, gender, birth date, occupation, mobile device brand, model, phone number, interest list and interest ranking). Once that the user is registered, the SWSR extracts the mobile user position (longitude and latitude data from the mobile device) every time when the user starts a session with the SWSR. Using all information, the SWSR creates an ontology instance of the user context, the mobile device and the user interests.

Using this application, any registered mobile user can search, select and invoke Web services according to his/her needs. Search of Web services starts when the user selects to view the list of Web services classified according to a given application domain. The Web service list is dynamically generated as a result of executing a query to the ontology model. The user has the option to search for Web services considering his geographical position at the time of the request. Once the user selects an application domain, the mobile application displays the interface, where the user can select from a list of recommend services based on the selected domain and user context. This recommendation is done through an inference rule, which is executed dynamically with user data. Let's assume that the user requested the address of any restaurant located at Lindavista district, as a result the mobile application returned a map location of a restaurant at Lindavista, recommended based on the users interests. The mobile application described in this work shows that the development of a multi-dimensional representation using ontologies and semantic Web services facilitates the generation of intelligent and dynamic recommendations of relevant services to end users. This result is mainly due to the incorporation of the user's context shaped by his interests and his geographical context.

This paper describes a multi-dimensional ontology model which represents three dimensions: user context information, Web services and application domain classifications. Using ontological representation enables semantic representation of concepts (classes), relations between concepts, individuals and inference rules to discover and produce new knowledge. In particular, the model reported in this paper allows the combination of different ontologies to offer a more complex representation of contextual variables and software resources. Using a multidimensional ontology modeling approach has the following benefits [6]:

a) Ontologies are managed as modules that can be expanded, reduced and maintained individually by their owners.
b) The multi-dimensional ontology model is itself another ontology, which imports ontologies as modules. In this multi-dimensional ontology model, an integrator defines semantic relationships across ontology modules regarding application interests.
c) Through the use of inference and query rules, the multi-dimensional ontology model can be used to answer questions traversing dimensions.

III. RELATED WORK

In the last years several projects concerning web services and mobile computing have been developed and reported. Of particular interest, are those related context-aware systems that incorporate multiple ontologies, use Web services for attending mobile user demands, incorporate a semantic Web service model and use reasoning and inference facilities for recommendations. This section presents an overview of these related work.

In 2004 Weißenberg et al. [7] presented FLAME2008, a platform for service customization through the use of information based on individual situations and personal demands of users. This proposal tries to determine the most appropriate set of web-based information and services through the semantic descriptions of situations and services. The main objective of FLAME2008 is to implement a Web-Based information system for large users groups and large service sets. In FLAME2008, services and information are sent to the mobile device, based on the current situation and profile of each user. The situation of user is obtained through the use of sensors and the information on user’s profile. Ontologies are responsible of matching demands and offers. Offers are composed of situation, profiles and bound services for the situations. Description of situations and services are based on profiles, they contain a set of attributes characterizing the situation. The values for the attributes are instantiated by the ontology. Sensor data and user’s profile are used to infer a demand. As a result, an ontology is created, where instances are used to construct a situation request profile semantically matched with all the situations known by the system. In this way a service request profile is constructed and matched against all the registered service profiles i.e. inferences are made on situations and services. Interests and preferences are specified in profiles. Besides of location, time and situation, the user’s context history is considered in the personalization of services and the information delivered to the mobile device. In FLAME2008 contexts are handled in a dynamic way, they are obtained from information of sensors and help to identify the situation of user; under this perspective the parameters of the situation define the context, giving the possibility of having a large range of contexts.

Ontologies in FLAME2008 are structured in layers: upper ontologies, which are used for processing generic and abstract concepts; domain ontologies, used to model concepts form different application domains; task ontologies, which model service ontology and situation ontology; and application ontology. In particular, the service ontology adhering to the OWL-S specification defines services using: profile for advertising and finding services, process model for describing cooperation between services and grounding for the execution of services. The situation ontology is defined by user's context and profile, this ontology is composed of several sub-ontologies for all the different context dimensions.

In 2004 Sheshagiri et al. [8] presented myCampus, a project where users subscribe with a set of task-specific agents to develop different tasks. These agents require the knowledge of contextual attributes of users. The sources of contextual information are modeled as Semantic Web Services that can be automatically discovered by agents. The static knowledge about users is stored in the form of rules that map contextual attributes onto service invocations, in this way the ontology can identify and activate the resources in response to the context of the users' queries. Sources of contextual information are defined as Semantic Web Services; it means that every source has a profile with the description of its functional properties. The ontology makes use of rules for local and global service identification, to identify one or more relevant sources of contextual information. Service discovery is carried out through these profiles.

FLAME2008 and MyCampus were two of the first projects that incorporated a semantic Web service approach by using OWL-S. However, the main drawback of OWL-S model is that it makes difficult the automatic incorporation of existing Web services (legacy WSDL files) requiring their semantic extension with OWL-S.

In 2005 Gu et al. [9] presented SOCAM, one of the first ontology-based models to represent contexts; SOCAM includes person, location, and activity and computer entity. SOCAM is a middleware architecture for rapidly building context-aware services. It provides support for discovering, acquiring, interpreting and accessing context information. Although SOCAM uses the concept of services, these are not defined by standardized languages such as WSDL or OWL-S. Limiting the possibility of incorporating existing Web services from different vendors on different platforms.

In 2005 Kim et al. [10] described a framework to search products and services through the use of a real-time ontology mapping mechanism between heterogeneous ontologies and taxonomies. This proposal is based on Web services and Semantic Web and is composed of: service client, service providers and search agent. The service client installed on a mobile device allows the specification of a product and a search intention. When a search is specified the information from de GPS is automatically considered. If a response is produced the client service is responsible for communicating results to the user. The service client uses a specific ontology to store user's specific categories and attributes related to ontologies. The service provider component allows the description of products and services through the providers' ontologies. When products and services are published providers must specify the meta-data of documents, they are needed by the ontology to find the description of services.

The main component of this proposal is the Service Search Agent, it is responsible of harvest the service providers through the use of a semantic web robot. The information collected is integrated into the ontology of products and services. The search agent receives the search request from client; it searches for relevant service providers, evaluates the obtained results and recommends the most adequate services for the user. The search agent is composed of four main components: an information integrator it uses a semantic robot for collecting the category and attribute information about products and services from various providers. When the request from client arrives the location processor defines the range of regions to be searched according to the location of user. Then with the information on the client's request and his location the query generator generates the query for the products and services ontology. The last component is the service evaluation agent that evaluates the retrieved results according to the intent on the user's search and recommends the most relevant products and services. This framework has a disadvantage, it requires services providers to describe their services using their own ontologies, leaving aside the use of interoperable Web service description languages. Another disadvantage is that it does not incorporate the reasoning services for deducing and generating recommendations.

In 2008 Dickson et al. [11] presented the use of Multi-Agent Systems, Semantic Web and Ontologies (called MAIS) for the implementation of an ubiquitous touristic service. Their objective was to provide coordination and integration of information and service resources anytime anywhere and provisioning personalized assistance to tourists. In the MAIS architecture tourist inquiries are sent to an ontology, results are produced according to the requirements and preferences of users. Through the use of the ontology the system can propose tour plans, formulate itinerary plans and connections between transport routes. The ontology organizes tourism-related information and concepts allowing the interoperability through the use of a shared vocabulary and meaning of terms. In this way all the agents in the system have a common basis for searching, interpreting and reasoning. The ontology was populated by extracting information and services from Web pages, using Web crawlers. Moreover this proposal has a Matchmaking mechanism to select and compose the packages that better correspond to the needs and preferences defined in the tourist's profile.

In 2009 Amel Bouzeghoub et al. [12] presented a context aware semantic recommender system. Recommendations are based on a multidimensional ontology which models persons, buildings, events and available resources. The recommender system was implemented for mobile users in a campus environment. Recommendations are proactively generated considering user context, geographic position and recommendation logs. Authors suggest that context is a multi-dimensional space, where each dimension is represented as a specific ontology. In particular, the set of ontologies incorporated into the multi-dimensional space are: domain ontology, user ontology, activity ontology, location ontology, and time ontology.

In 2009 Cadenas et al. [13] described mIO, an ontology network to represent knowledge related to context. The mIO ontology consists of a core ontology which interlinks different ontology modules needed for modeling context. They introduce the concept of modularization for ontologies, to allow using only the modules that are involved in a given case, instead of using the whole ontology network. Their ontology network contains ten modular ontologies: User, Role, Environment, Location, Time, Service, Provider, Device, Interface and Network. Authors also propose using context logs for better recommendation results.

Despite the use of an ontological model in Dickson [11], Bouzeghoub [12], and Cadenas [13] projects, their main drawback is that they do not incorporate semantic Web service representation and reasoning.

In 2009 Sousa [14] presented ICAS, an architecture that allows creation of context aware services, and SeCoM, a semantic model to represent contexts. Authors present a study case applied to a university campus, where the objective is to identify pedagogic characteristics of documents and persons. SeCoM model consists of six main ontologies and six supporting ontologies. Main ontologies are: Actor ontology, Time ontology, Temporal Event ontology, Space ontology, Spatial Event ontology, Device ontology, and Activity ontology. Secondary ontologies are: Contact ontology, Relationship ontology, Role ontology, Project ontology, Document ontology and Knowledge ontology. Inference is based on user preferences, user context, and user interests to identify relevant services. In this approach, services are classified in categories, adding preconditions and post conditions for each service, allowing search and selection of services by comparing input and output compatibilities. This work also maintains action logs, which are used for recommendation.

In 2009, Woerndl and Hristov [15] described an approach for personal information management in mobile devices, using a recommender system based on ontologies. The system recommends documents and articles considering time and location context and the user personal ontology. Authors implemented a PDA Semantic Desktop (SeMoDesk). Recommendations are obtained from the interest of a user in a topic or document, considering user schedule and location. SeMoDesk is a desktop application for mobile devices, for this reason has some limitations regarding the ontology model and does not support semantic Web service invocations.

In 2009 Liiv [16] describe SMARTMUSEUM, a platform for recommendations in the cultural domain of a museum. SMARTMUSEUM uses a combined approach based on rules, collaboration and content personalization, where content is semantically enabled by an ontology. Recommendations are about cultural objects allocated in the museum and content related with those objects. User profile includes abilities and interests of user, and a log of visited places.

In 2010, Gómez-Pérez presented SEEMP [17] a project aiming at facilitating employment services in Europe. SEEMP consists of an ontology network which describes employs and employees from human resource perspective. SEEMP reference ontology consists of the following ontologies: Job Seeker Ontology, Job Offer Ontology, Compensation Ontology, Driving License Ontology, Economic Activity Ontology, Occupation Ontology, Education Ontology, Geography Ontology, Labour Regulatory Ontology, Language Ontology, Skill Ontology, Competence Ontology, and Time Ontology.

These reported proposals are based on similar technological mechanisms, such as context-awareness, incorporating multiple ontologies (multi-dimensional space), use Web services for attending mobile user demands, incorporate a semantic Web service model and use reasoning and inference facilities for recommendations. However, the main difference between related work and the approach described in this paper, is the implementation of a multi-dimensional ontology model with adaptable and extendible ontology modules; and the incorporation of a semantic Web service representation capable of acquiring legacy WSDL files. As a result, the approach reported in this paper offers an innovative contribution for dynamic and changing mobile environments.

IV. Multi-dimensional Ontology Model

The core solution of this work consists of a multi-dimensional ontology model, for which the following design objectives were established:

a) Build a model capable of representing multiple dimensions with changing attributes.
b) Design the model using a semantic formalism which allows the description of classes (concepts), class hierarchies, semantic relationships between those concepts, and axioms.
c) Design and implement dimensions as self-contained ontologies to enable modularization. Modularization of ontologies in turn facilitate individual ontology maintenance, update and expansion.
d) Design the model to allow the integration of multi-dimensional ontology models to face and solve multi-disciplinary problems.

e) Enable the definition of query functions to extract information using any number of dimensions and any number of attributes.

f) Enable the definition of inference rules which allow the generation of new semantic connections between concepts across all dimensions.

g) Enable automated inclusion of pure WSDL service descriptions into a ontological representation of services. Without imposing Web service providers a new requirement for augmenting their services with very specific models such as OWL-S.

In order to achieve the afore-mentioned design objectives, a multi-dimensional space model was implemented with ontologies. OWL [18] was chosen as the ontological language, because it is based on description logics (DL) allowing the description of concepts and semantic relations between concepts. For inference rules definition, SWRL was selected, because it is fully compatible and importable into OWL ontologies, so through a set SWRL rules new semantic relations can be deduced logically. To query the model, SQWRL [19] was used. SQWRL (**S**emantic **Q**uery-enhanced **W**eb **R**ule **L**anguage) is built based on the well known SWRL which allow extensions by built-ins. SQWRL defines a set of built-ins operators that can be used to construct more specialized functions for querying ontologies. The multi-dimensional ontology model consists of three dimensions: the *user context*, the *application domain* and the set of available *services*; each of these dimensions define multiple and changing attributes. For instance, to model the interest of the user, it is necessary to consider a wide range of possibilities, depending on the subject of interest.

The architecture depicted in Figure 1 shows the multi-dimensional ontology model and mobile applications which exploit the information modeled and represented in the ontological model. Representing multiple dimensions of semantic information using Web-based ontologies is a promising trend from the area of knowledge representation that has proven good results. An important benefit of using multi-dimensional ontologies is the feasibility of maintaining each ontology and the possibility of exchanging and extending parts of the model.

In the following sub-sections each ontology is briefly described:

A. *User Context Ontology*

The user can define his context through the use of his mobile device. First the user must define his profile; the application for user's profile definition has been implemented with the *Framework JSF* and the *PrimeFaces* components. The first step in the profile definition is to specify the name, birth date, sex, and occupation of user. This information it is stocked into the users context ontology (Figure 2).

After the user supplies the information for profile creation a connection with the users context ontology is established and a primary list of interests is generated for the user. In this list the user can select the services where he is interested. Every item in the list has a level of interest property, which is used to assign a weight to the item. This weight means the level of interest of user on that service and he can change it through the mobile device.

The user context ontology represents the semantic information of the user context, incorporating his general data, occupation, interests and information from the mobile device used to interact with the system. In particular, the information required is related to its geographical position. Figure 2 shows classes of the *User Context* ontology, which are described next.

Interest, this class defines a hierarchy of concepts of interest to the user; *Interest Record*, this class represents the interaction between users and their reported interests (*interest level* defines a property that that takes values in a range from 1 to 10, indicating the level of interest that a user has in a given period of time over a specific concept), when the individuals for the *Interest* class are created they are assigned with an interest level of five, that is an intermediate value (see *inference rule* 1) that can be changed by the user. *User*, describes the general user characteristics represented by user name, date of birth and gender; *Occupation*, occupation defines a job or profession of the user, *Device*, describes the characteristics of the mobile device of the user; *Position*, defines the geographic coordinates of latitude and longitude obtained through the mobile device.

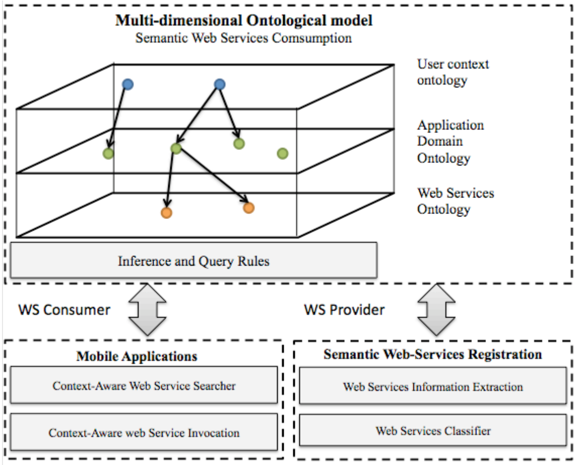

Multi-dimensional Ontological model
Semantic Web Services Comsumption
User context ontology
Application Domain Ontology
Web Services Ontology
Inference and Query Rules
WS Consumer
WS Provider
Mobile Applications
Context-Aware Web Service Searcher
Context-Aware web Service Invocation
Semantic Web-Services Registration
Web Services Information Extraction
Web Services Classifier

Figure 1. General architecture

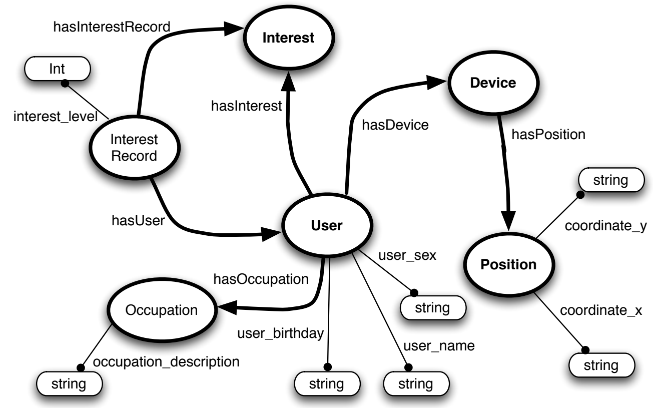


Figure 2. User Context Ontology

B. *Web Service Ontology*

The Web services ontology shown in Figure 3 shows the common components that any Web service describes. This ontology is populated automatically when a service provider is registered in the system and publishes Web service interfaces using WSDL files. This ontology allows Web services to be semantically annotated with more functional information. This ontology consists of the following classes: *Provider*, defines the individual Web services published by supplier name, password, email and a URL; *Service* defines a Web service provider entered by using a service name and a access URL, *Type*, defines complex data types used within the service using a type name, a base class and a boolean flag that determines if a data type is comparable with geographical longitude and latitude, *Operations*, defines the operations defined in the service description; *Variable* defines the input and output values of an operation, in addition to describing the components of a complex data type; *Value* samples, defines a set of values that can be used as a reference for assigning value a given input variable.

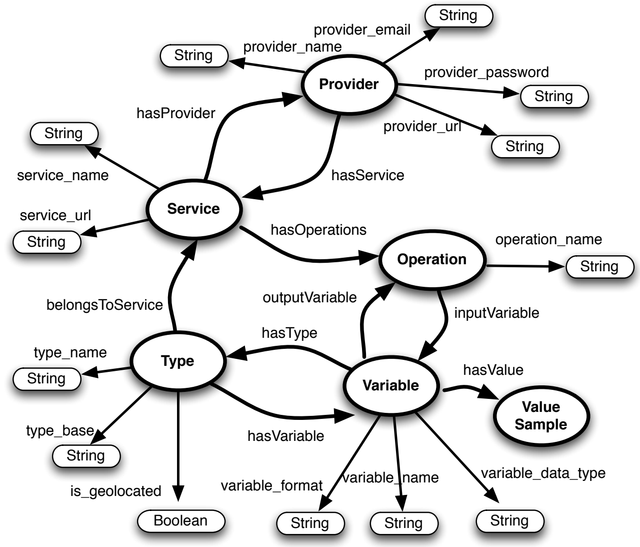


Figure 3. Web Service Ontology

C. *Application Domain Ontology*

The application domain ontology (Figure 4) defines a class hierarchy for classifying Web services according with a taxonomy of concepts related to various domains of interest to the user and applications that consume Web services. Through this ontology it is possible to find intersections between services functionalities and users' interests. This ontology can be continuously updating and adapting to new user requirements and new service providers offers.

The application domain ontology defines the class *Domain*, that defines a classification of possible fields of application of Web services and user interests.

D. *Integrating the Multi-dimensional Ontology*

This ontology imports all the definitions, concepts and semantic relations from the user context ontology, the Web service ontology and the application domain ontology.

In order to semantically relate the three models and produce new knowledge from them, it is important to establish semantic relationships between the concepts. However, these relationships are not defined arbitrarily, relations are decided based on a particular intention. In this case, the objective of putting the three ontological models together is to find related information with the user domain of interest and the Web service application domain. Therefore, the following relations are defined into the multi-dimensional ontology model:

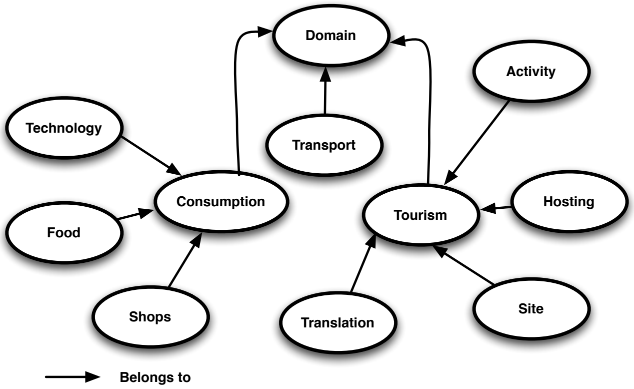


Figure 4. Application Domain Ontology

Based on the user context ontology internal relation *hasOccupation* (meaning: a user has a job or profession) the external relation *hasLevel* (meaning: an occupation has socioeconomical level) links the *Occupation* class with *Level* class from the application domain ontology. The semantic relation *interestHasDomain* (meaning: interest has an application domain) between the *Interest* class from the user context ontology and the *Domain* class from the application domain ontology correlates the user interest with application domains.

The semantic relation *serviceHasLevel* (meaning: a Web service has a socioeconomical level) correlates the *Service* class from the Web service ontology with the with *Level* class from the application domain ontology. And the semantic relation *serviceHasDomain* (meaning: a Web service has an application domain) correlates the *Service* class from the Web service ontology with the with *Domain* class from the application domain ontology, enabling with these relations to annotate semantically Web service definitions. Annotated Web services facilitate other service-related tasks such as Web service discovery and Web service matchmaking.

Finally, among an important semantic relation is the *userHasRecommendation* (meaning: a user has a context-based recommendation to consume a specific Web service). This relation enables the final user to get recommendations based on his/her interests and context. Figure 5 shows all semantic relationships between the three ontological models.

E. Inference and Query Rules

In order to discover and produce new semantic relations between individuals from the multi-dimensional ontology population, a set of inference rules are defined and executed.

Inference Rule 1. If a user *u*, is related to an interest *i*, through a record of interest *x*, and also the interest *i* has an application domain *d*, and service *s* also has the same domain *d* and the interest level *n* of the user *u* is greater than 5; then the inference engine makes the recommendation of the service *s* to the user *u*.

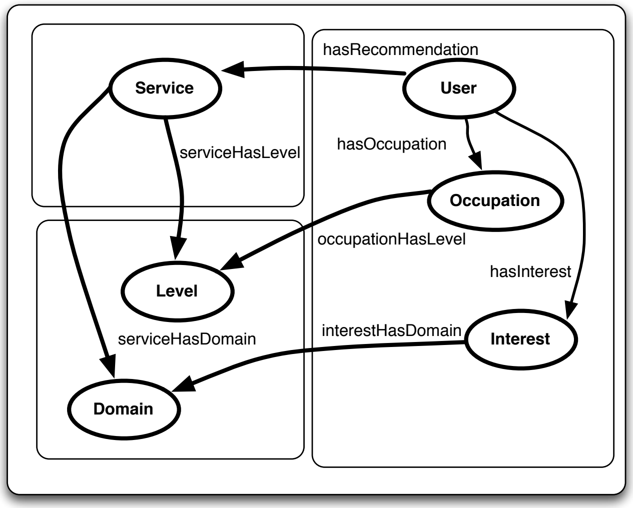


Figure 5. Multi-dimensional Ontology Model

contexto:Usuario(?u) ^ contexto:Intereses(?i) ^
dominios:Dominio(?d) ^ servicios:Servicio(?s) ^
contexto:RegistroInteres(?x) ^ contexto:tieneUsuario(?x, ?u) ^
contexto:tieneInteresesRegistrados(?x, ?i) ^ (1)
interesTieneDominio(?i, ?d) ^ servicioTieneDominio(?s, ?d) ^
contexto:nivel_interes(?x, ?n) ^ swrlb:greaterThan(?n, 5) ->
tieneRecomendacion(?u, ?s)

To facilitate external applications to query and search over the concepts and relations, a set of query rules are also defined and included into this multi-dimensional ontology.

Query Rule 2. This rule allows to search for services related with a specific Web service provider.

dominios:Dominio(?d) ^ servicios:Servicio(?s) ^
servicios:tieneProveedor(?s, ?prov) ^
servicioTieneDominio(?s, ?d) ^
servicios:nombre_servicio(?s, ?nombre) ^ (2)
servicios:url_servicio(?s, ?url) ->
sqwrl:select(?s, ?nombre, ?url, ?d)

Query Rule 3. This rule allows obtaining information about a particular user device.

contexto:Dispositivo(?dispositivo) ^
contexto:tieneDispositivo(?u, ?dispositivo) ^
contexto:marca_dispositivo(?dispositivo, ?marca) ^
contexto:modelo_dispositivo(?dispositivo, ?modelo) ^ (3)
contexto:numero_telefono_dispositivo(?dispositivo, ?tel) ->
sqwrl:select(?dispositivo, ?marca, ?modelo, ?tel)

Query Rule 4. This rule obtains the interest level of a given user, with respect to the interest defined in the ontology.

contexto:RegistroInteres(?r) ^
contexto:tieneInteresesRegistrados(?r, ?i) ^
contexto:tieneUsuario(?r, ?u) ^ (4)
contexto:nivel_interes(?r, ?nivel) ->
sqwrl:select(?r, ?i, ?nivel)

This multi-dimensional ontology can be enhanced by defining more inference and query rules to extract interesting information across dimensions.

Reasoning performance depends on the number of axioms and population of ontologies. In particular, in the multi-dimensional model reported in this paper, each ontology is maintained consistent by manually running checks periodically. So far, the number of individuals and axioms in ontologies remain low, so performance problems with reasoning tasks have not been faced. However, it is highly likely that when the number of Web services grows scaling problems will arise. To cope with this problem, there is the plan to manage interchangeable service ontologies.

In the case of rule-based reasoning, until now, no performance problems have been faced, this is mainly because the model uses few inference rules, most of the rules are query-rules, which consume less resources.

V. WEB SERVICE SEMANTIC REGISTRY APPLICATION

In order to populate, modify and query the multi-dimensional ontology model described in Section III, a *Web Service Semantic Registry* (WSSR) was implemented (see Figure 6).

Figure 6 shows the use case diagram of the WSSR application. This application interacts with two main actors: the Web service provider and the ontology manager. The former is the person responsible for registering provider data and Web services. The ontology manager is a sub-system which offers administrative functions, such as: automatic Web service parsing, to extract relevant Web service data; automatic Web service recording into the ontology; automatic link between application domains and context; and execution of inference rules to produce recommendations for Web service consumption.

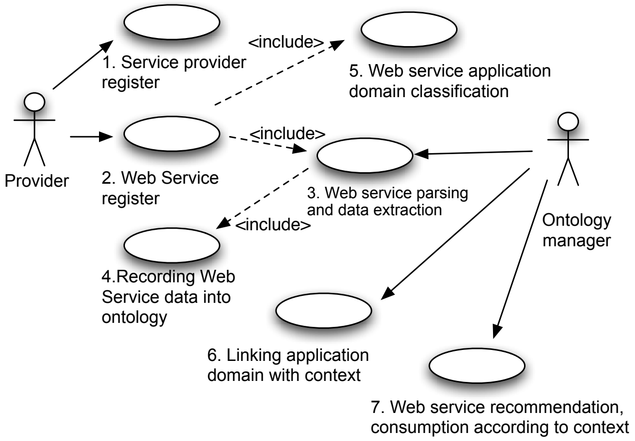

1. Service provider register
<include>
5. Web service application domain classification
Provider
2. Web Service register
<include>
3. Web service parsing and data extraction
Ontology manager
<include>
4.Recording Web Service data into ontology
6. Linking application domain with context
7. Web service recommendation, consumption according to context

Figure 6. Use cases and associations of WSSR application

The WSSR was implemented as a Web application to enable Web service providers around the world to access the system and perform any of the following tasks: registration of new Web service providers, publication of Web services, semantic classification of Web services, and Web service lists. In this section these functionalities are described.

A. Registration of Web Service Providers

To gain access to the WSSR and make use of the registry functions, service providers must be registered first. Every new service provider is required to enter the following data: full name, URL of provider´s website and email address. He is also asked to create a username and password to access his account. All this information is stored in the multi-dimensional ontology model. In this study case, a total of 45 different Web service providers were registered in the ontology. In turn, each of these providers published different Web services.

B. Registering Web Services

The application for Web services registration, requires the provider to enter the URL of the WSDL file. The application connects to the specified URL, parsers the WSDL file to extract the relevant Web service information.

Registering Web services is an automated procedure which is executed through the following tasks:

1. *Data types information extraction*, this task consists of retrieving XML-Schemas from the WSDL file to obtain detailed information of complex and simple data types defined in the Web service interface.
2. *Extraction of service name and operations*, this task gets the name attribute of the service tag and the names of service operations. It also reads the elements that describe input and output parameters of each service operation.
3. *Registration of Web service information* in the web services ontology, this task creates a new individual instance in the *Service* class, using the specified URL and service name as data type attributes. It also establishes a semantic relation between the registered service and its corresponding provider using the *hasProvider* and *hasService* relations.
4. *Registration of parameters individuals* for complex data types into the *Parameter* class, establishing a new semantic relation between the new service instance and its parameters.
5. *Registration of operations individuals* in the class *Operation*, establishing respective semantic links with input and output parameters and parameters samples. Furthermore entity relationships are created among those individuals. There is a particular case when the return data type of a service operation is classified as geographically locatable. With this data type, it is possible to display the results on a map.

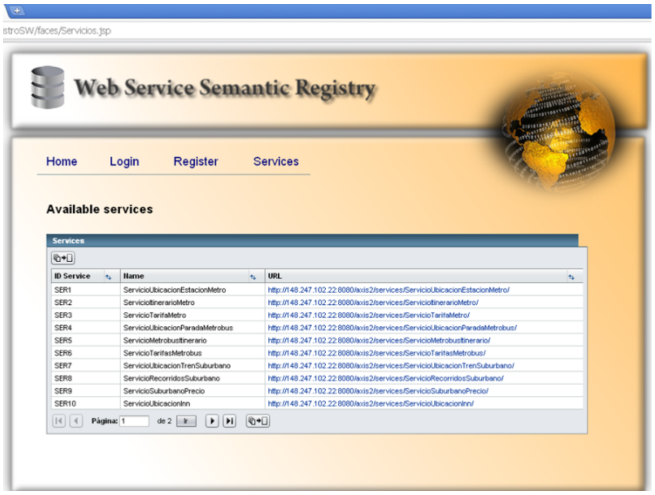
Using this facility, a set of 20 Web services were registered in the WSSR. Figure 7 shows the Web interface listing deployed Web services. These Web services are related with transportation, hotel reservation, flight booking, tours, and general offers. Table 2 shows a classification of implemented Web services.

Table 1 Implemented Web Services

Transport	Hotel Booking	Flight Booking	Tours	Restaurant
Metro	Fiesta Inn	Aviacsa	Travelocity	Sanborns
Metrobús	Fiesta Americana	Aeroméxico	Turissste	Chili´s
Tren suburbano	Crowne Plaza	Volaris	Turística 2000	Sam´s Club
	Emporio	Inter Jet		Toks
	Sheraton			Steren

C. Web Services Classification

Once the provider has registered his Web services (one at a time), he must classify them, this is done by using the classification tool of the WSSR application, where the provider of web services has to select an application domain, a sub-domain and a type from the domain ontology to. It is also necessary to define a socioeconomic level, depending on the target market where the provider wants to offer his Web services. Classification of services is extremely important because if a service is not classified, it will not be possible to link it with any consumer request. It is also important to note that this classification can be extended or even exchanged with another application domain or domains if required. Currently domain-based classification of Web services is done semi-automatically. However, one of future extensions of this work is to implement fully automated classification of published Web services.



istroSW/faces/Servicios.jsp

Web Service Semantic Registry

Home Login Register Services

Available services

Services

ID Service	Name	URL
SER1	ServicioUbicacionEstacionMetro	http://148.247.102.22:8080/axis2/services/ServicioUbicacionEstacionMetro/
SER2	ServicioItinerarioMetro	http://148.247.102.22:8080/axis2/services/ServicioItinerarioMetro/
SER3	ServicioTarifaMetro	http://148.247.102.22:8080/axis2/services/ServicioTarifaMetro/
SER4	ServicioUbicacionParadaMetrobus	http://148.247.102.22:8080/axis2/services/ServicioUbicacionParadaMetrobus/
SER5	ServicioMetrobusItinerario	http://148.247.102.22:8080/axis2/services/ServicioMetrobusItinerario/
SER6	ServicioTarifasMetrobus	http://148.247.102.22:8080/axis2/services/ServicioTarifasMetrobus/
SER7	ServicioUbicacionTrenSuburbano	http://148.247.102.22:8080/axis2/services/ServicioUbicacionTrenSuburbano/
SER8	ServicioRecorridosSuburbano	http://148.247.102.22:8080/axis2/services/ServicioRecorridosSuburbano/
SER9	ServicioSuburbanoPrecio	http://148.247.102.22:8080/axis2/services/ServicioSuburbanoPrecio/
SER10	ServicioUbicacionInn	http://148.247.102.22:8080/axis2/services/ServicioUbicacionInn/

Página: 1 de 2 Ir

Figure 7. Web service semantic registry showing deployed services

VI. CONTEXT-AWARE MOBILE APPLICATION

To evaluate the proposed model, a mobile application was implemented. This mobile application allows any user to register into the WSSR to search for web services and invoke them as requested. This section describes the mobile application modules implemented.

A. *Mobile user registration*

All new mobile users are requested to register into the system before making use of its facilities. This registration task is executed as follows:

a) *Obtaining mobile user data*, this task requires the user to interact with the mobile application to enter the following data: full name, gender, birth date, occupation, mobile device (brand, model and phone number), interest list and interest ranking.

b) *Extracting mobile user position*, this task is executed as many times as the user logins the system through the mobile application. Therefore, position is the dynamic value which encapsulates longitude and latitude data from the mobile device.

c) *Registration of mobile user data and context*, this task registers a new user individual into the user context ontology. Also a new individual is created into the ontology to represent the mobile device characteristics. Occupation is established as a semantic relation between the user individual and the occupation class. The list of interests is also recorded in the ontology with their respective semantic associations with the user. Figure 8 shows the graphical user interface of the mobile application.

B. *Web Services Search and Invocation*

Search, selection and consumption of Web services are among the most demanded application tasks related with Web services. This functionality was implemented in the mobile application (see Figure 9). Search of Web services starts when the user selects to view the list of Web services classified according to their application domain. The classified Web service list is dynamically generated as a result of executing a query to the ontology model. The user has the option to search for Web services considering his geographical position at the time of the request.

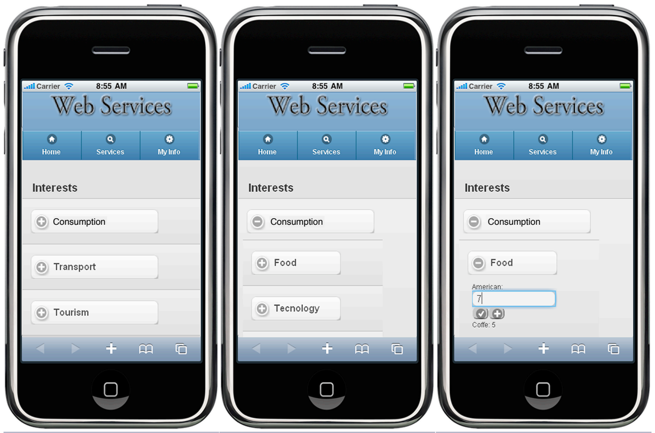
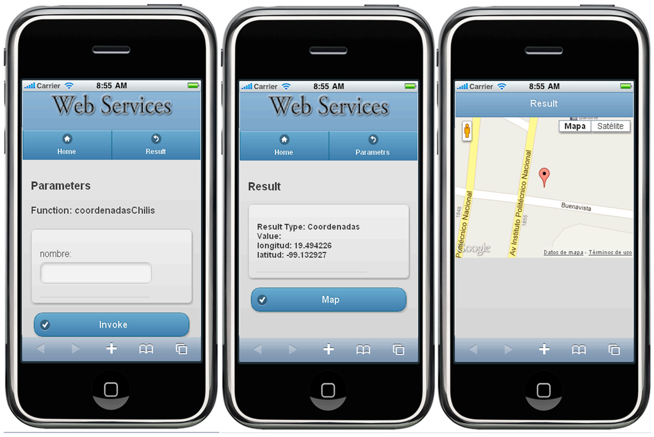


Figure 8. User registers interests in the mobile application

Once the user selects an application domain, it displays the interface, where the user can select from a list of recommend services based on the selected domain and user context. This recommendation is done through an inference rule, which is executed dynamically with user data.



Function: coordenadasChilis

nombre:

Result Type: Coordenadas
Value:
longitud: 19.494226
latitud: -99.132927

Figure 9. Search, selection and invocation of Web Services

Figure 9 shows the result of invocating a location Web service. In particular, in this case the user requested the address of any restaurant located at Lindavista district, as a result the mobile application returned a map location of a restaurant at Lindavista, recommended based on the users interests.

The mobile application described in this work shows that the development of a multi-dimensional representation using ontologies and semantic Web services facilitates the generation of intelligent and dynamic recommendations of relevant services to end users. This result is mainly due to the incorporation of the user's context shaped by his interests and his geographical context.

VII. Conclusions

This paper describes a multi-dimensional ontology model which incorporates the user context information, semantic Web services interface modeling and application domain classifications. The work reported in this paper incorporates various technological paradigms, such as: semantic Web services, mobile computing and ontologies. The main objective of integrating these technologies was to support the development of more complex and intelligent mobile context-aware applications.

In this paper we reported the implementation of a mobile-based architecture for the use of Semantic Web Services trough a multi-dimensional ontology. We present the initial results. Performance results are going to be obtained once the automatic mechanism for populating the ontologies from web pages is finished.

The use of multi-dimensional models implemented with ontologies offers significant advantages: the ability to exchange, expand, extend and maintain the individual ontologies. An example is the application domain ontology, which can be interchanged as needed to adapt to new application needs.

The incorporation and exploitation of Web services through ontological models is a clear trend that promises to improve the automatic selection and invocation of legacy and new Web services.

All these technologies together (Web services and ontologies) are key facilitators for the wise management of context-based systems based on mobile computing.

Acknowledgment

This work has been supported by CONACyT, México.

References

[1] Y. Park and F. Adachi (eds.), Enhanced Radio Access Technologies for Next Generation Mobile Communication, pp. 1–37. 2007 Springer.

[2] Mobile Web Initiative. The Web and Mobile Devices. 2010. http://www.w3.org/Mobile/. March 2012

[3] K. Chen and J. de Marca, Mobile Wimax Wiley 2008.

[4] M. Weiser ”The Computer for the 21st Century”, ACM Mobile Computing and Communication Review, vol. 3, issue 3, pp. 3-11, 1999, doi: 10.1145/329124.329126

[5] N. Blefari-Melazzi, E. Casalicchio, and S. Salsano, “Context-aware Service Discovery in Mobile Heterogeneous Enviroments”, Proc. IEEE Mobile and Wireless Communications Summit, 2007, 16th IST, septembre 2007 pp. 1-5, Budapest, doi: 10.1109/ISTMWC.2007.4299226

[6] M. Horridge, H. Knublauch, A. Rector, R. Stevens, and Ch. Wro, A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools, Edition 1.0, The University Of Manchester, March 2012

[7] N. Weißenberg, A. Voisard, and R. Gartmann “Using Ontologies in Personalized Mobile Applications”, Proceedings of the 12th annual ACM international workshop on Geographic Information Systems (GIS 04), November 2004, pp. 2-11, doi: 10.1145/1032222.1032225

[8] M. Sheshagiri, N. Sadeh, and F. Gandon “Using Semantic Web Services for Context-Aware Applications”, Proceedings of Mobisys2004 Workshop on Context Awareness, June 2004.

[9] T. Gu, H. Pung, and D. Zhang, "A service-oriented middleware for building context-aware services", Journal of Network and Computer Applications, vol. 28, issue 1, january 2005, pp. 1-18, doi: 10.1016/j.jnca.2004.06.002.

[10] W. Kim, S. Lee, and D. Choi “Semantic Web Based Intelligent Product and Service Search Framework for Location-Based Services”, Lecture Notes in Computer Science Springer, Computational Science and Its Applications (ICCSA 2005) Part IV, may 2005, pp. 103-112, Singapore, doi: 10.1007/11424925_13

[11] K. Dickson, W. Chiu, T. Yves, F. Yueh, L. Ho-fung. C. Patrick, and K. Hung “Towards ubiquitous tourist service coordination and process integration: A collaborative travel agent system architecture with semantic web services” ACM Information Systems Frontiers, Volume 11, Number 3, July 2008, pp. 241-256, Bradford, UK., doi: 10.1007/s10796-008-9087-2

[12] A. Bouzeghoub, K. Ngoc, and L. Krug, "Situation-Aware Adaptive Recommendation to Assist Mobile Users in a Campus Environment", Proceedings IEEE International Conference on Advanced Information Networking and Applications (AINA-09), May 2009, pp. 503-509, doi: /10.1109/AINA.2009.120

[13] A. Cadenas, C. Ruiz, I. Larizgoitia, R. Garcia, C. Lamsfus, I. Vázquez, M. González, D. Martín, and M. Poveda, “Context Management in Mobile Environments: a Semantic Approach” ACM Proceedings of 1st Workshop on Context, Information and Ontologies, CIAO 2009, June 2009, Heraklion, Grecia, doi: 10.1145/1552262.1552264

[14] P. Sousa, E. Carrapatoso, B. Fonseca, M. Campos, and R. Bulcão-Neto, “Composition of context aware mobile services using a semantic context model”, International Journal On Advances in Software, volume 2, numbers 2 and 3, December 2009, pp. 275-287.

[15] W. Woerndl and A. Hristov, “Recommending Resources in Mobile Personal Information Management”, IEEE Proceedings of the Third International Conference on Digital Society, (ICDS 2009), February 2009, pp. 149-154, doi: 10.1109/ICDS.2009.21.

[16] I. Liiv, T. Tammet, T. Ruotsalo, and A. Kuusik, "Personalized Context-Aware Recommendations in SMARTMUSEUM: Combining Semantics with Statistics," IEEE Procedings of Third International Conference on Advances in Semantic, October 2009, pp. 50-55, 2009, doi: 10.1109/SEMAPRO.2009.25.

[17] M.C. Suárez-Figueroa, A. Gómez-Pérez, and B. Villazón-Terrazas, “How to Write and Use the Ontology Requirements Specification Document”, ACM Procedings Confederated International Conferences, CoopIS, DOA, IS, and ODBASE 2009 on On the Move to Meaningful Internet Systems: Part II OTM’09, November 1009, pp. 966-982, 10.1007/978-3-642-05151-7_16

[18] S. Bechhofer, M. Dean, F. Van Harmelen, J. Hendler, I. Horrocks, D. McGuinness, P. Patel-Schneider, G. Schreiber, and L. Stein,: OWL Web Ontology Language Reference, W3C Proposed Recommendation. http://www.w3.org/TR/owl-ref/. March,2012.

[19] Semantic Query-Enhanced Web Rule Language. http://protege.cim3.net/cgibin/wiki.pl?SQWRL March 2012

About an Architecture That Allows to Become a Mobile Web Service Provider

Marc Jansen
Computer Science Institute
University of Applied Sciences Ruhr West
Bottrop, Germany
marc.jansen@hs-ruhrwest.de

***Abstract*—The role of mobile devices as Web Service consumers is widely accepted and a large number of mobile applications already consume Web Services in order to fulfill their task. Nevertheless, no reasonable approach exists, as yet, to allow deploying Web Services on mobile devices and thus uses these kinds of devices as Web Service providers. This paper presents an approach that allows deploying Web Services on mobile devices by the usage of the well-known protocols and standards and, at the same time, can overcome problems that usually occur when mobile devices are used as service providers. Here, we provide both the description of an implementation with results of a first performance test. The test shows that the described approach provides a reasonable way to introduce Web Service provisioning for mobile devices.**

Keywords - mobile devices; Web Services; mobile Web Service provider.

I. Introduction

In recent years, the number of reasonably powerful mobile devices has much increased. According to [1], the number of smartphones worldwide counts about 300 million units.

On the other hand, this huge number of smartphones represents a large number of heterogeneous devices with respect to the operating systems smartphones are currently using. According to [2], there were at least five different operating systems for smartphones available on the market in 2010, and their distribution is shown in Fig. 1.

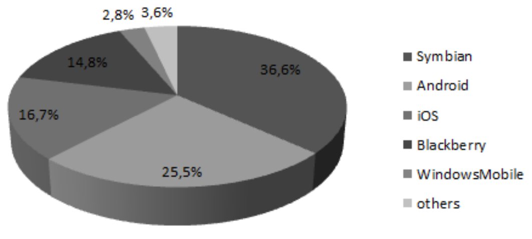


Figure 1. Distribution of different operating systems for smartphones in 2010

It thus seems to be necessary to have a platform-independent mechanism for the communication with services provided by smartphones in order to not re-implement each service for each of the mentioned operating systems.

Usually, Web Services are used in order to provide a standardized and widely used methodology that is capable of achieving a platform-independent way to provide services. Unfortunately, in contrast to consuming Web Services on mobile devices, providing Web Services on mobile devices is not yet standardized due to several problems that occur when a service runs on a mobile device.

This paper presents the description of a framework that allows providing Web Services on mobile devices. The outline of the paper is as follows: the next section provides an overview of related work, after which the scenario - together with the problems that usually occur should Web Services be provided by a mobile device - is explained. The following section explains the implementation of the framework in detail and the results of a first performance test are presented. The paper is closed by a conclusion.

II. State of the Art

The idea of providing Web Services on mobile devices was probably presented first by IBM [3]. This work presents a solution for a specific scenario where Web Services are hosted on mobile devices. More general approaches for providing Web Services on mobile devices are presented in [4] and [5]. In [6], another approach, focusing on the optimization of the HTTP protocol for mobile Web Services provisioning, is presented.

Importantly, none of the mentioned approaches manages to overcome certain limitations of mobile devices, as demonstrated in the next section.

The major difference between previous research and the approach presented in this paper is that, to the best of our knowledge, previous research focused very much on bringing Web Services to mobile devices by implementing server side functionality to the mobile device in question. The approach presented here follows a different line: from a technical and communication point of view, the mobile Web Service provider communicates as a Web Service client with a dynamically generated Web Service proxy.

This approach provides an advantage for overcoming certain problems with mobile Web Services as described in the next section. Furthermore, this approach does not rely on an efficient server side implementation of Web Services on the mobile device, and thus allows to implement a very lightweight substitution to a common application server where a common Web Service is running.

Since nothing comes for free, this approach has some drawbacks as well, e.g., it implements a polling mechanism that permanently polls for new service requests. Therefore, this approach produces an overhead with respect to the network communication and the computational power of the mobile device. The computational overhead, though, can be dramatically reduced by adjusting the priority of the polling mechanism according to the priority of the provided Web Service.

Another drawback of the presented approach is that it relies on a publicly available proxy infrastructure for the part of the framework that dynamically generates the Web Service proxies. This drawback can be overcome if, for example, mobile telecommunication companies provide this kind of infrastructure centrally.

In contrast to the before mentioned approaches, the approach presented in this paper differs with respect to one major aspect: from a network technical point of view there is no server instance installed on the mobile device. Therefore, a certain Web Service client does not call the Web Service on the mobile device directly but calls a centrally deployed proxy. The Web Service running on the mobile device polls in regular intervals for any new message requests of interest. The sequence of the Web Service request from the client point of view and from the Web Service point of view is shown in the sequence diagram in Fig. 2.

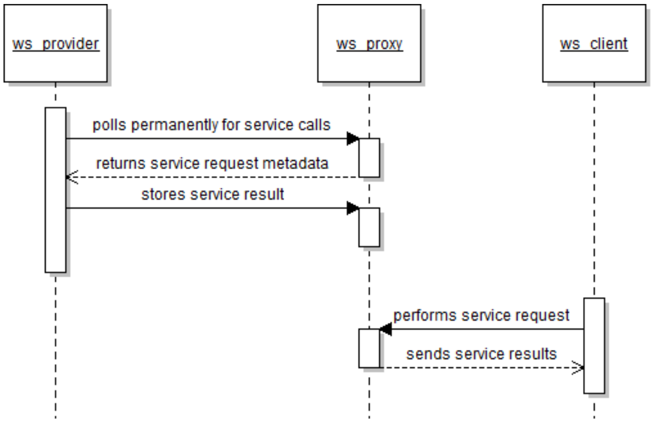


Figure 1: Sequence diagram of the Web Service calls in the presented approach

The exact sequence of the different messages and events will be described in more detail later. Since especially polling mechanisms cause a certain drawback, one of the major questions concerning the presented approach is the question of benefits and drawbacks of the polling mechanism and, in particular, whether the benefits justify the drawbacks.

As already mentioned, one of the major problems of dealing with Web Services on mobile devices is the fact that mobile devices often switch between networks. Therefore, the Web Service running on a mobile device is usually not available under a fixed address, a fact that leads to a number of problems for the consumer of a mobile Web Service: Besides the usual network switch, the fact that mobile devices are usually not meant to provide 24/7 availability, but are designed towards providing the user with the possibility to exploit certain services, e.g., phone calls, short messages, writing and receiving emails, etc., yields the problem that mobile devices might get switched off by the user. Hence, not only that the provided Web Service might be unavailable under different network addresses, but it might not be available at all.

All these drawbacks can be solved by using the approach presented here. By using the central proxy, the service requests of a certain Web Service client can be stored and if the mobile Web Service is running, it can pull for service requests that are of interest to it. Since from a technical point of view the Web Service provider only acts as a client to the Web Service proxy, the potentially changing network addresses of the mobile device do not pose a problem at all.

In addition, one of the major drawbacks of the described polling mechanism can be limited by adjusting the priority of the Web Service running on the mobile device, resulting in a lower frequency of the polling for the service request.

To conclude, in our opinion, the advantages of the described mechanism justify the drawbacks that are inherent to the approach.

III. Scenario Description

The major idea behind the implementation of the middleware is to provide a Web Service proxy, according to the proxy design pattern [7], in order to overcome certain problems in mobile scenarios as described by [8]. One major problem here is that mobile devices often switch networks, e.g., at home the mobile phone might be connected to a WiFi network, at work the connection might be established through another WiFi network and on the way home from work the mobile phone might be connected to a GPRS/UMTS-network. Each of these different networks provides different IP addresses and possibly different network constellations. For example, it can be private IP addresses with network address translation (NAT), where the Web Services running on the device are not directly accessible from the internet, or public IP addresses.

Frequently switching between IP addresses might raise certain problems for the provision of Web Services, since the client of a certain service always needs to know the actual IP address at which the service can be reached. More than that, within a private network the provided Web Services are usually not reachable at all from the internet.

Therefore, the problem, from the client point of view, is that the service is not always accessible under the same (and constant) IP address. The presented approach provides a solution to overcome this problem, with the exception of the case when a device is completely switched off. The switch off problem can be overcome as well, in which case slight modifications to the presented approach, together with an asynchronous call of the Web Service, are necessary.

The approach presented here suggests solving these problems by implementing a Web Service proxy that dynamically creates a proxy for each Web Service that gets deployed on a mobile device. The created proxy allows receiving service requests as a representative to the actual service and storing a service request along with the necessary data. In the next step, the mobile Web Service provider continuously polls for requests to its services, performs the services and sends the result back to the dynamically generated Web Service proxy. Receiving the result, the Web Service proxy can send the result back to the client that originally performed the service request.

IV. Implementation

The major goal of the work presented here is to provide a solution to the described scenario. Therefore, we implemented a middleware that allows the provision of Web Services on mobile devices. Here, the standard protocols, e.g., WSDL for the description of the Web Service interface, SOAP/REST as the standard network protocol and http as the usual transport protocol, are used such that there is no additional effort on the client side for requesting a mobile Web Service.

The following three sections provide a short introduction to the services offered by the middleware, followed by a description of the communication between the mobile Web Service provider and the Web Service client/consumer. Last but not least some details are presented about the Java based implementation for the test scenario.

A. Use-Case Analysis

In order to achieve the goal of implementing a Web Service proxy, an analysis of use-cases that this proxy will have to support has been performed. The result of this analysis is shown in Fig. 3.

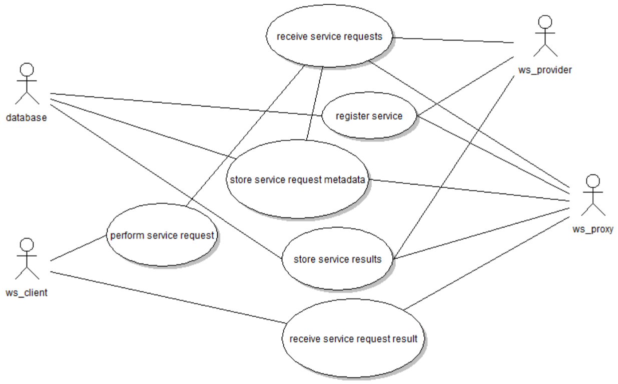


Figure 2: Use case description of the developed middleware.

From a technology point of view four different actors participate in the scenario. Obviously, a provider for the mobile Web Service is necessary. This is a piece of software running on the mobile device that provides the Web Service itself. This piece of software can best be compared with an application server hosting a Web Service in a scenario where the Web Service is provided by a common server system.

The second quite obvious actor is the consumer of the Web Service: the Web Service client. This is a piece of software running on the client side, performing requests to the Web Service.

As already described, one of the major ideas of the presented approach is to provide a proxy for the Web Services provided by the mobile devices. Therefore, the Web Service proxy is another actor that participates in the scenario. The proxy represents a surrogate of the Web Service provided by the mobile device. The basic function of this proxy is to implement the same interface (same methods with identical parameter lists and return values) as the Web Service itself. Moreover, the methods provided by the proxy (in order to register a service, de-register a service, etc.), should be accessible via the standard network protocols of Web Services and the description of the proxy interface should also be available in WSDL (in the implementation here the SOAP protocol was chosen). The proxy's' major task is to receive client requests, store them in a database and wait for the mobile Web Service to provide the result of the service request. While in the traditional proxy pattern the proxy would directly forward (push) the incoming service requests to the Web Service, we have decided to just store the requests in a database in order to allow the mobile Web Services to pull the requests from the proxy. This change to the traditional proxy pattern basically allows handling constantly changing network connections (as explained before), since within this approach neither the Web Service proxy nor the Web Service client need to know the actual IP address of the mobile device that provides the actual Web Service.

Fourth and last, the database is taken to be an actor of the middleware. Usually, the database would more likely be modeled as a system (and not as an actor), but for the sake of clarity and consistency, we decided to model the database also as an actor in the system. The major task of the database is to store the necessary information about the service request in order to allow the Web Service running on the mobile device to perform the requested task, and to later-on store the return values of the service request as well. By storing also the return value, the Web Service proxy is able to send the result back to the client that made the request. This is necessary since the usage of the proxy is transparent to the client, in the sense that the client is not aware that the actual service request is not answered by the proxy, but by the Web Service running on the mobile device. Therefore, the Web Service proxy needs to send the result of the service to the Web Service client, and not the mobile Web Service itself.

Besides the four actors, a number of use-cases need to be implemented in order to fully run the described scenario:

First of all, a mobile Web Service provider needs to be able to register a service to be provided. Besides the Web Service provider, the Web Service proxy and the database are interacting within this use-case, too. The Web Service proxy needs to dynamically implement the interface of the mobile Web Service and the storage of the metadata (basically the name of the method that should be called and its parameter values) of the service requests. The database needs to provide certain storage for the parameter values of each method (in case of a relational database: a table) and the according return values of the mobile Web Service.

The second, quite obvious, use-case is that the mobile Web Service provider needs to be able to receive service requests. Besides the mobile Web Service provider, the Web Service proxy participates in this use-case also, since this is the instance that directly receives the requests from the Web Service client and stores the necessary information in the database. Two additional use-cases, namely, perform service requests and receive service request results, participate in the store service request metadata use-case.

Additionally, we have identified two other use-cases that are necessary for the handling of the service request metadata (store service request metadata) and the handling of the return values (store service result). The first of these two use-cases interacts with two actors: the Web Service proxy and the database; the second one additionally interacts with the Web Service Provider.

Beside the fact that the provision of these use-cases allows the implementation of the described scenario, one of the major advantages of this approach is that the Web Service client only interacts with the preformed service request and receives corresponding answers from the service request result use-case. Therefore, from a client point of view, the request to a mobile Web Service is no more than a usual service request. No additional effort is necessary on the client side in order to receive results from a Web Service running on a mobile device.

B. Communication between the mobile Web Service and its clients

In order to explain the necessary communication for a service request from the Web Service client to the mobile Web Service provider, we modeled the communication flow within the sequence diagram shown in Fig. 4.

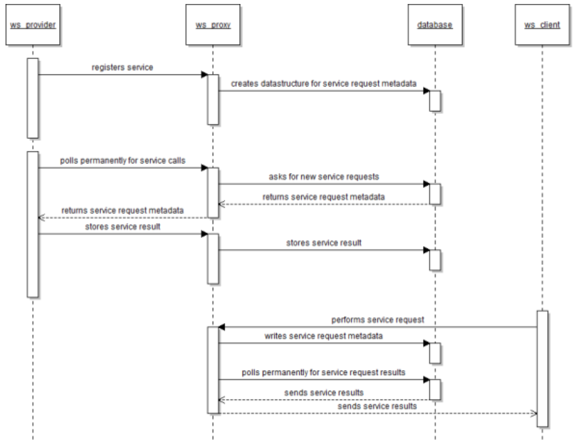

ws_provider
ws_proxy
database
ws_client
registers service
creates datastructure for service request metadata
polls permanently for service calls
asks for new service requests
returns service request metadata
returns service request metadata
stores service result
stores service result
performs service request
writes service request metadata
polls permanently for service request results
sends service results
sends service results

Figure 3: The UML sequence diagram for the communication between a Web Service provider and its client.

Within the sequence diagram we have modeled an object life line for each of the actors, to be discussed later. First of all, the mobile Web Service provider needs to register its service with the Web Service proxy. As part of the service registration process the Web Service proxy creates the necessary data structure for storing the service requests in the database.

After the mobile Web Service provider has registered its service, it permanently polls the Web Service proxy for new service requests. The Web Service proxy asks the database if a new service request for the respective mobile Web Service provider is available and if so, returns the request's metadata to the mobile Web Service provider. After receiving the metadata of a new service request, the mobile Web Service provider performs the service and sends the result of the service to the Web Service proxy that directly stores the result in the database.

From a client point of view, the Web Service client simply calls the service provided by the Web Service proxy. While receiving a new service request, the Web Service proxy stores the necessary request metadata in the database. Afterwards the Web Service proxy directly starts to permanently poll the database for the result of the respective service request. Once the mobile Web Service provider has finished performing the request and has stored the result (via the Web Service proxy) in the database, the Web Service provider is able to send the result of the service request back to the client.

C. A sample implementation

In order to test the described approach with respect to its performance, we implemented the Web Service proxy in Java. Additionally, the mobile Web Service provider was implemented for Android. Here, we focused on an intuitive and easy way for the implementation of the Web Service, and have therefore, oriented ourselves by the JAX-WS (Java API for XML-Based Web Services), as described in the Java Specification Request 224 (JSR 224). The major idea, adapted from JAX-WS, was that a Web Service can easily be implemented by the use of two different annotations: the @MobileWebService annotation marks a class as a Web Service, and methods within this class can be marked as methods available through the mobile Web Service with the @MobileWebMethod annotation.

With the help of these two annotations a simple mobile Web Service, which only calculates given integer values, can be implemented as follows:

```
@MobileWebService
public class TestService {

       @MobileWebMethod
       public int add(int a, int b) {
              return a + b;
       }

}
```

The basic relationships between the major classes of the sample implementation are shown in Fig. 5. For the sake of simplicity and transparency, less important classes (and methods of each class) have not been modeled.

Basically, the implementation consists of two packages. Package one is the proxy package which is usually deployed on a server that is reachable from the internet via a public IP address. Here, we find one class that implements the necessary methods for the registration of a new mobile Web Service, the permanent polling from the mobile Web Service for the service request metadata and the method that allows storing the result of the service request in the database. All these methods are reachable as Web Services themselves, so that the communication between the instance running the mobile Web Service and the Web Service proxy is completely Web Service-based.

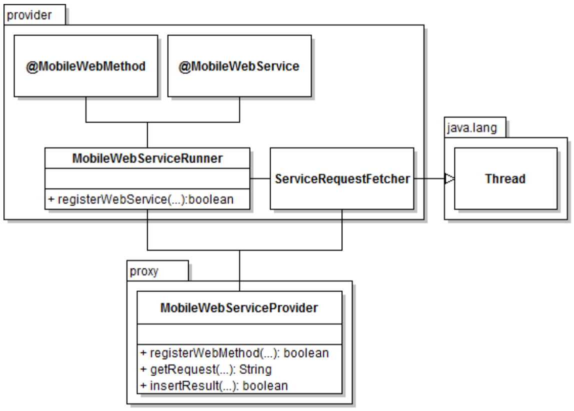


Figure 4: UML class diagram of major parts of the sample implementation

Basically, the implementation consists of two packages. One package that is usually deployed on a server that is reachable from the internet via a public IP address, this is the proxy package. Here, we find one class that implements the necessary methods for the registration of a new mobile Web Service, the permanent polling from the mobile Web Service for the service request metadata and the method that allows to store the result of the service request in the database. All of these methods are themselves reachable as Web Services, so that the communication between the instance running the mobile Web Service and the Web Service proxy is also completely Web Service based.

In the provider package we find, as one of the major classes, the MobileWebServiceRunner class to which the mobile Web Service gets deployed. This class is basically comparable to an application server in a common Web Service environment, but with a dramatically lower footprint. This lower footprint is extremely important to mobile devices due to their usually limited resources. Additionally, this package also provides the two formerly mentioned annotations that allow an easy marking of a class as a mobile Web Service and, accordingly, a certain method of such a class as a mobile Web Method. Last but not least, this package also implements the ServiceRequestFetcher class. This class inherits the java.lang.Thread class since its responsibility is to permanently poll the Web Service provider for new service requests.

V. Performance tests

Since the communication is a little bit more complicated, in comparison to a common Web Service call, one concern of this approach is the question of its performance. In order to get a first idea of how good or bad this implementation behaves with respect to performance issues, we implemented a simple performance test.

A. Description of the test scenario

For the performance test we implemented a very simple mobile Web Service. This service only calculates the sum of two given integers and returns the respective value as the result. The major advantage of such a simple mobile Web Service is that almost the entire duration of the mobile Web Service call is dedicated to the communication, and almost no amount of the round-trip time is used for the calculation itself. Since the communication is the complex part of the presented approach, we assume that this method of performance testing would provide the best overview about the communication performance of the presented approach. In the test scenario a common client (running on a common PC) had to put a number of service requests to the mobile Web Service.

In order to compare the results against the performance of common Web Service calls, we implemented the test scenario also the other way around: we implemented a common Web Service (running on a common server) and called this Web Service from a mobile device. Here, the basic idea was to use the same hard- and software-environment with minimal changes and also to maintain the same network environments in all of the tests.

In addition, we were interested in the communication performance in different network settings. Therefore, we performed the same tests in four different network settings. For each of the tests the (mobile) Web Service and its consumer where running:

- … in the same (WiFi) network,
- … different networks, and the mobile device was connected via WiFi,
- … different networks, and the mobile device was connected via UMTS
- … different networks, and the mobile device was connected via GPRS

We conducted eight different test cases: four for the different network constellations with a mobile Web Service running on a mobile device and a Web Service client running on a common PC, and four test cases where the Web Service was running on a common Server and the client was running on a mobile device.

In the test cases where the (mobile) Web Service provider and the client were not connected to the same network, the central components have been deployed to a server running via Amazon Web Services (AWS), as a Cloud Computing provider.

B. Test results

Within each of these eight test cases, one hundred service calls were performed and the duration of each call was measured.

The results for the mobile Web Service in the different network scenarios are shown in Fig. 6.

As expected, the performance for the mobile Web Service calls was pretty good and pretty constant in the case the mobile device was connected with a WiFi network. If both the mobile Web Service provider and the client were connected to the same WiFi network, the average duration was M = 147.69ms (SD = 76.00ms). Having the mobile Web Service provider connected to a different WiFi network, the average duration for one service call was M = 339.04ms (SD = 61.71ms).

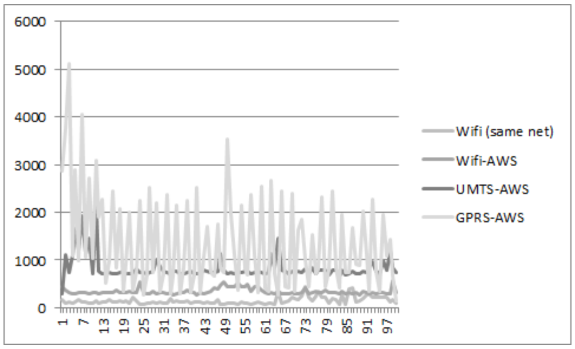


Figure 5: Results for the mobile Web Service in the different network constellations

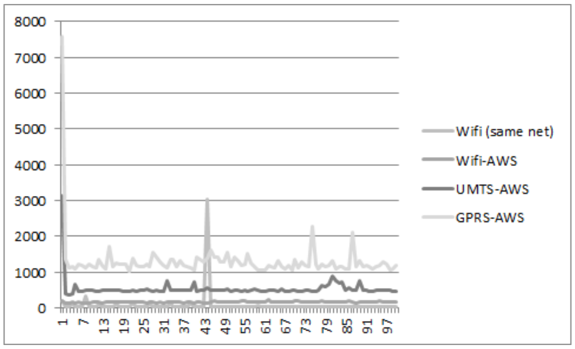


Figure 6: Results for the usual Web Service calls in the different network constellations

As expected the performance for the mobile Web Service calls are pretty good and pretty constant if the mobile device is connected with a WiFi network. The average time if both the mobile Web Service provider and the client are connected to the same WiFi network was M = 147.69ms (SD = 76.00ms). Having the mobile Web Service provider connected to a different, still WiFi, network the average time for one service call calculates to M = 339.04ms (SD = 61.71ms).

Of course, we measured less performance of the service calls when the mobile Web Service provider was connected to a mobile network, the performance of the service calls was lower. The results for the UMTS based network connection of the mobile Web Service show an average of M = 827.55ms (SD = 250.35ms) for each service call, while the results for the GPRS based network are even worse. Here, the average for a single service call is M = 1355.96ms (SD = 986.38ms). As can be seen from the values for the standard deviation, the performance of single service calls differs dramatically as well, e.g., the minimum duration measured within the UMTS scenario was MIN = 283ms and the maximum was MAX = 2169ms. The results for the GPRS based scenario are even worse, with a MIN = 142ms and MAX = 5123ms.

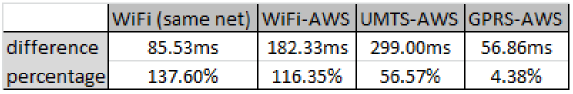
The task of the second step of the test was to compare the performance results with the performance of a common Web Service call. For that purpose we conducted the same test, but this time the Web Service was not running on a mobile device but on a common server, while the Web Service client was running on a mobile device - again in the four different network settings. The results of these tests are shown in Fig. 7.

As demonstrated, the results are better from both perspectives - the overall performance and the standard deviation in the different network settings. A common Web Service call, if the Web Service provider and the mobile Web Service consumer are connected to the same WiFi network, has an average round-trip duration of M = 61.16ms (SD = 301.36ms). When the Web Service client was connected to a different (still WiFi) network the average performance was M = 156.71ms (SD = 15.24ms).

Here, again, the values for the Web Service client connected to a mobile network are somewhat lower. In the case of the UMTS network, the average service call showed a performance of M = 528.55ms (SD = 273.34ms), and the results for the GPRS based network even worse with an average for each of the service calls of M = 1299.10ms (SD = 658.75ms).

The next step was to compare the different results. The major goal of this comparison was to get an idea of how good the performance of the presented approach for mobile Web Service calls is, in comparison to common Web Service calls. Therefore, we calculated the difference in the average performance of a single Web Service call in the different scenarios first, and as a second step calculated the percentage of the performance difference in the different scenarios. The results are shown in Table 1.

TABLE 1: COMPARISON OF THE COMMON WEB SERVICE CALLS AND THE MOBILE WEB SERVICE CALLS IN THE DIFFERENT NETWORK SCENARIOS



	WiFi (same net)	WiFi-AWS	UMTS-AWS	GPRS-AWS
difference	85.53ms	182.33ms	299.00ms	56.86ms
percentage	137.60%	116.35%	56.57%	4.38%

The table shows that, in comparison to common Web Service calls, the performance of the presented approach was not too good when the mobile Web Service was connected to a WiFi network. The results for the mobile Web Service provider and the client connected to the same network showed a performance overhead of 137.60 per cent, and when the mobile Web Service was provided within a different WiFi network the performance overhead was about 116.35 per cent. But, if the mobile Web Service was connected to a mobile network, the performance overhead was not that dramatic anymore. In the case of the UMTS network the overhead was limited to 56.57 per cent, and for the GPRS based network the overhead was even lower at 4.38 per cent. Therefore, on the basis of our test results, it can be said that the performance of the presented approach for mobile Web Services (in comparison to common Web Services) seems to improve the lower the network bandwidth is. This could best be seen by the results for the GPRS based network, where the actual overhead in our test was below 5 per cent.

VI. Conclusions

As demonstrated in this paper, today's modern and powerful mobile devices can be used as Web Service providers by using well-known and accepted standards and protocols. The presented approach is capable of solving some of the problems that usually occur while providing Web Services on mobile devices, e.g., the problem of constantly changing IP addresses. Furthermore, the overhead that is inherent in the presented approach does not seem to be a show stopper. As shown, the performance in commonly available mobile networks, like UMTS or GPRS, is comparable to common Web Service calls.

It can, therefore, be concluded that the presented approach provides an interesting alternative to the common Web Service provisioning by using mobile devices that act as a server also from a technical point of view. It eliminates certain problems that usually occur if mobile devices provide Web Service provider infrastructures, and the resulting drawbacks from the performance point of view are acceptable.

Having in mind the power that the presented approach would provide for new approaches and scenarios, it could be asserted that bringing Web Services to mobile devices will probably become more important in the future and that we will most likely see an increasing number of applications making use of that kind of technology.

Acknowledgment

This work was partly supported by an Amazon AWS research grant.

References

[1] IDC Worldwide Quarterly Mobile Phone Tracker, January 27, 2011.

[2] Tudor, B. and Pettey, C., 2010. Gartner Says Worldwide Mobile Phone Sales Grew 35 Percent in Third Quarter 2010, Smartphone Sales Increased 96 Percent, Gartner, http://www.gartner.com/it/page.jsp?id=1466313, last visited 19.11.2011

[3] McFaddin, S., Narayanaswami, C., and Raghunath, M., 2003. Web Services on Mobile Devices – Implementation and Experience, In: Proceedings of the 5th IEEE Workshop on Mobile Computing Systems & Applications (WMCSA'03), pp. 100-109, Monterey, CA

[4] Srirama, S., Jarke, M., and Prinz, W., 2006. Mobile Web Service Provisioning, In: Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT/ICIW 2006), p. 120, Guadeloupe, French Caribbean

[5] AlShahwan, F. and Moessner, K., 2010. Providing SOAP Web Services and REST Web Services from Mobile Hosts, In: Fifth International Conference on Internet and Web Applications and Services (ICIW), pp. 174-179.

[6] Li, L. and Chou, W., 2011. COFOCUS – Compact and Expanded Restful Services for Mobile Environments, In: Proceedings of the 7th International Conference on Web Information Systems and Technologies, pp. 51-60, Noordwijkerhout, The Netherlands

[7] Gamma, E., Helm, R., Johnson, R., and Vlissides, J., 1995. Design Pattern – Elements of Reusable Object-Oriented Software, pp. 185-195, Addison-Wesley.

[8] Svensson, D., 2009. Assemblies of Pervasive Services. Dept. of Computer Science, Institutional Repository – Lund University.

Extending OWL-S for the Composition of Web Services Generated With a Legacy Application Wrapper

Bacem Wali
INSERM/INRIA/Univ. Rennes 1, VISAGES U746
Faculté de médecine
Rennes, France
bacem.wali@irisa.fr

Bernard Gibaud
INSERM/INRIA/Univ. Rennes 1, VISAGES U746
Faculté de médecine
Rennes, France
bernard.gibaud@irisa.fr

Abstract— **Despite numerous efforts by various developers, web service composition is still a difficult problem to tackle. Lot of progressive research has been made on the development of suitable standards. These researches help to alleviate and overcome some of the web services composition issues. However, the legacy application wrappers generate nonstandard WSDL which hinder the progress. Indeed, in addition to their lack of semantics, WSDLs have sometimes different shapes because they are adapted to circumvent some technical implementation aspect. In this paper, we propose a method for the semi automatic composition of web services in the context of the NeuroLOG project. In this project the reuse of processing tools relies on a legacy application wrapper called jGASW. The paper describes the extensions to OWL-S in order to introduce and enable the composition of web services generated using the jGASW wrapper and also to implement consistency checks regarding these services.**

Keywords- Ontology; Semantic Web; Web Services Composition

I. INTRODUCTION

Web services are a new revolution of software systems. They are considered as self-contained, self-describing, module applications that can be published, located, and invoked through the Web [1][2]. They are designed to be manipulated remotely from a network and they have the capability to invoke each other mutually, which raises the issue of their interoperability. Companies implement web services according to their application domain and display them through the web. Consequently, the number of heterogeneous web services is increasing, whose interoperability is severely hampered by this pervasive heterogeneity, inherent to independently developed services. For example: The use of new messaging protocols involves changing WSDL formats according to the domain specific applications and implementation needs. Therefore, when we deviate from standard cases to specific ones, composition of web service becomes a challenging problem that was addressed by many researchers and engineers in the recent years [3].

Different initiatives have been proposed to facilitate the reuse of web services, leading to new languages, protocols and frameworks. For example, UDDI [4] (Universal Description, Discovery and integration), SOA (Service Oriented architecture) [5], BPEL4WS (Business Process Execution Language for Web Service) [6], SOAP [7] (Simple Object Access Protocol) and WSDL (Web Services Description Language) [8] are standards for service discovery, description, and messaging protocols [1][9]. Those specifications provide a means to syntactically describe a web service. However, they do not deal with semantic web service description and semantic web service composition.

Semantic web service is a concept that brings semantics to the aforementioned standards. By adding semantics we can make web services machine understandable and use-apparent form [10]. By adding semantic markup to a web service we can make two aspects of its functionality explicit. First, semantic annotations can define what the service actually does, and second, they can describe its behavioral aspect (i.e., how the service works, the chaining that can be performed according to the sent and received messages).

Several languages have emerged, to add semantic description features to the web services standards. For example, DAML-S [11] (Darpa Agent Markup Language for services) is a revision of DAML+OIL [12] and based on OWL [13] (Ontology Web Language) WSMO [14] (Web Service Modeling Ontology) and OWL-S (Ontology Web Language for Services) [15]. OWL-S is an ontology represented in OWL which aims at applying reasoning capabilities to the functionality, behavior, and execution of web services. OWL-S defines a model with three layers: a **service profile** describing the service's basic functionalities (function and characteristics, etc), a **service model** describing how the service works including data and control constructs flow, and a **service grounding**, describing how to access the service, by grounding its functional elements (input, outputs and operations) in a way consistent with the WSDL's concept of binding.

We worked in the context of the NeuroLOG [16] project which aims to share medical resources (brain images and image processing tools) [17]. Image processing tools are wrapped as web services using a software package called jGASW [18]. JGASW is a framework for wrapping legacy scientific applications as web services enabling their execution in a SOA environment.

To allow the sharing of neuro-imaging resources, the OntoNeuroLOG [19] ontology was designed. It provides common semantics for information sharing throughout the NeuroLOG system and allows the sharing of neuro-imaging resources provided by collaborating actors in the field of neuro-imaging research**.** The term resources cover both neuro-imaging data (such as images) as well as image processing tools (registration, de-noising, and segmentation).

Through the semantic we tend to share the functionalities of images and the functionalities of image processing tools to enable more expressivity from a functional viewpoint.

The goal of our work is to facilitate the sharing, reuse, and invocation for the user of image processing tools wrapped as web services with jGASW and deployed in the site server within the NeuroLOG framework. To this end, we chose to add semantics to jGASW services, to facilitate workflow composition and automate some consistency controls regarding their usage. Our assumption is that this may increase the usability of such tools by people that were not involved in their development. This by addressing some technical aspects that hinder the composition process with OWL-S API and implementing some consistency controls. Such consistency control tends to: (i) ensure interoperability and composition by checking the compatibility between outputs and inputs of web services; (ii) check of the compatibility between the inputs provided by the users and the semantic inputs definition of the service; (iii) check the consistency between the functionality of the image processing tool, like registration or de-noising, and their declared inputs/outputs, with respect to the formal definition of such conceptual actions, modeled in the OntoNeuroLOG ontology.

We used the OWL-S ontology to benefit from its web service description model and its large expressivity in terms of parameters description and behavioral aspect of flows. However, we had to deal with some technical issues regarding jGASW WSDLs which let us extending OWL-S to enable its use in our NeuroLOG framework.

In this paper, we advance the state of the art by (1) specifying an extension of the OWL-S specification to make it adapted to our jGASW framework context without changing the basic structure of the WSDLs, (2) adding some reasoning capabilities to perform consistency checks regarding the usage of our annotated web services. The following of the paper is organized as follows: Section II provides more details about the difficulties related to the WSDLs of service generated by the jGASW tool, together with OWL-S semantic descriptions of services. Section III presents the solutions that we found based on a specific extension of OWL-S that addresses the problem and some associated reasoning mechanisms that validate the services capabilities, consistent with our domain ontology OntoNeuroLOG. Section IV details how the implementation was done to extend OWL-S and solve the problem and describes some technical issues that we tackle. Section V discusses our contribution and situates it in the wider context of semantic workflows, and finally, Section VI opens other perspectives for future work.

II. BACKGROUND

To address the issue of web service composition using OWL-S and jGASW we need to describe both more closely. So, we first describe WSDL files generated automatically by jGASW, then we study the automatic generation of semantic descriptions with OWL-S and we analyze the mismatch between the two and we study how to address it.

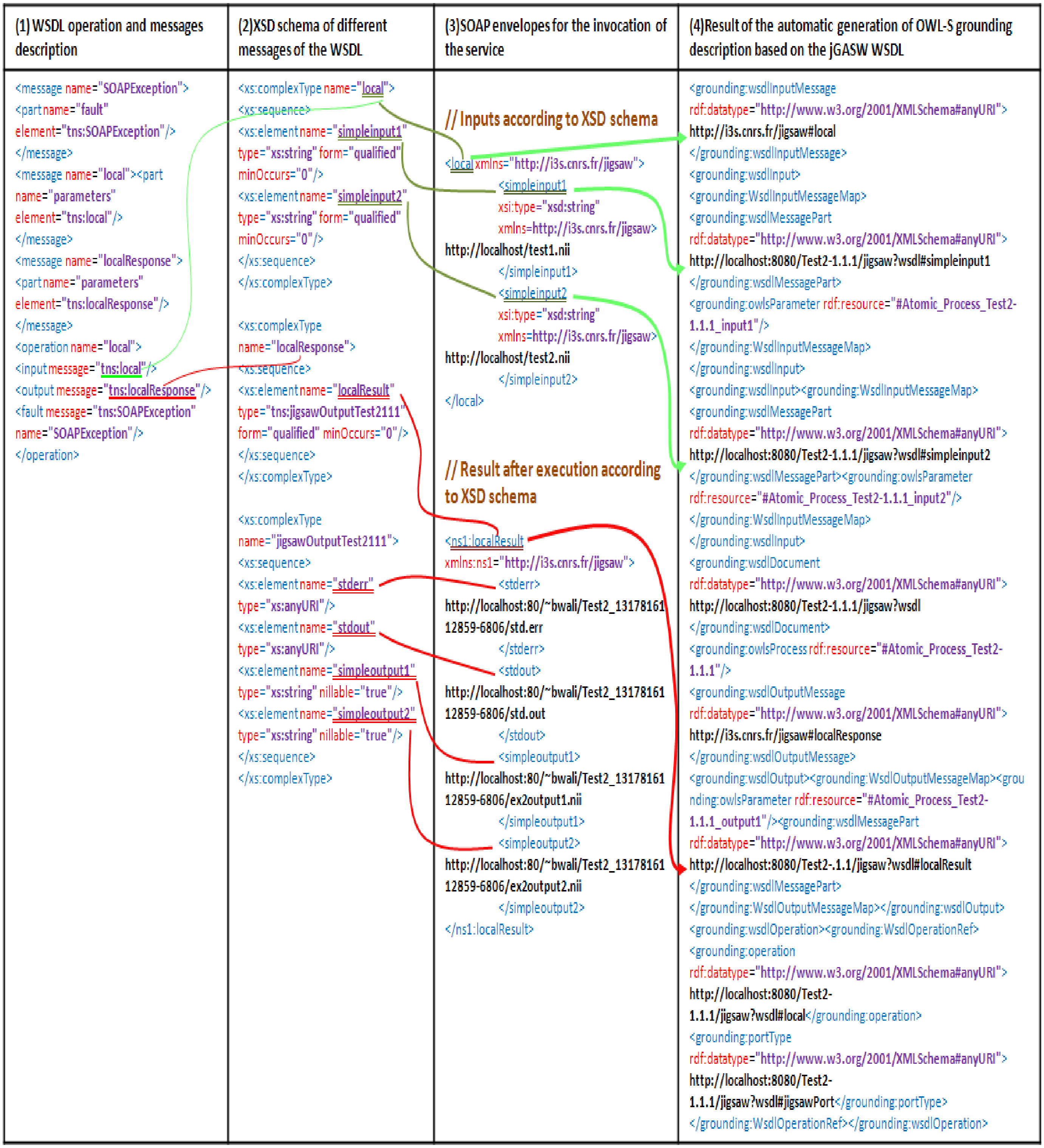
First, we explain how the jGASW [18] framework works: an application GUI allows the user to upload the image processing tools (shell program) and add inputs, outputs arguments and libraries. According to an XML schema and values set by the user, an XML description is generated *jGASW descriptor*. The generation of the web service consists in transforming the *jGASW descriptor* into a web service interface by generating the WSDL together with an XSD schema (XML Schema Definition). The XSD schema details the inputs and the outputs of every WSDL operation. In fact, all services have WSDL files with the same content, and identical operations but always have only one input and one output and one output as an exception. In contrast, inputs and outputs are described differently in the XSD schema according to each service. For example, in the Figure 1, column 1 illustrates the description of a WSDL operation named *local* that is composed of *tns:local* as input, *tns:localResponse* as output, and *tns:SOAPException* as fault message (generated if the execution of the service failed). Column 2 details input *tns:local* by defining it as a **complexType containing two *xs:element*** (i) (simpleinput1 and simpleinput2: input files) and details output *tns:localResponse* by defining it as a **complexType typed as another complexType** (ii) (jigsawOutputTest2111: generated automatically). This complexType contains four ***xs:element*** representing the different output files (stdout, stderr, simpleoutput1, simpleoutput2).

Column 3 shows SOAP envelopes (call/request). At execution time, jGASW prepares the SOAP envelope, invokes the service and gets back the result according to the sequences described in the complex type jigsawOutputTest2111, but the name of the envelope is *ns1:localResult*. Thus, the jigaswOutputTest211 is not considered here.

JGASW wraps executables into web services, and produces at least two standard files (std.out: for standard output for any shell output, std.err: error message generated (if execution fails), and other files resulting of the service execution such as image files (e.g. ex2output1.nii, ex2output2.nii).

Column 4 in the figure 1 shows the OWL-S description generated automatically from the WSDL description provided by the jGASW service (using the WSDL2OWLS [20] converter). Green arrows show the grounding of each individual input element whereas the red arrow shows one unique grounding which is *localResut*. In fact, here, we lost the information that this output contains four files rather one only. Thus, composition of jGASW services is not possible.

So, the principal issue concerns the outputs definition, which is not understandable due to their complex schema.



(1) WSDL operation and messages description

```
<message name="SOAPException">
<part name="fault"
element="tns:SOAPException"/>
</message>
<message name="local"><part
name="parameters"
element="tns:local"/>
</message>
<message name="localResponse">
<part name="parameters"
element="tns:localResponse"/>
</message>
<operation name="local">
<input message="tns:local"/>
<output message="tns:localResponse"/>
<fault message="tns:SOAPException"
name="SOAPException"/>
</operation>
```

(2)XSD schema of different messages of the WSDL

```
<xs:complexType name="local">
<xs:sequence>
<xs:element name="simpleinput1"
type="xs:string" form="qualified"
minOccurs="0"/>
<xs:element name="simpleinput2"
type="xs:string" form="qualified"
minOccurs="0"/>
</xs:sequence>
</xs:complexType>

<xs:complexType
name="localResponse">
<xs:sequence>
<xs:element name="localResult"
type="tns:jigsawOutputTest2111"
form="qualified" minOccurs="0"/>
</xs:sequence>
</xs:complexType>

<xs:complexType
name="jigsawOutputTest2111">
<xs:sequence>
<xs:element name="stderr"
type="xs:anyURI"/>
<xs:element name="stdout"
type="xs:anyURI"/>
<xs:element name="simpleoutput1"
type="xs:string" nillable="true"/>
<xs:element name="simpleoutput2"
type="xs:string" nillable="true"/>
</xs:sequence>
</xs:complexType>
```

(3)SOAP envelopes for the invocation of the service

```
// Inputs according to XSD schema

<local xmlns="http://i3s.cnrs.fr/jigsaw">
          <simpleinput1
          xsi:type="xsd:string"
          xmlns=http://i3s.cnrs.fr/jigsaw>
http://localhost/test1.nii
          </simpleinput1>
          <simpleinput2
          xsi:type="xsd:string"
          xmlns=http://i3s.cnrs.fr/jigsaw>
http://localhost/test2.nii
          </simpleinput2>
</local>

// Result after execution according
to XSD schema

<ns1:localResult
xmlns:ns1="http://i3s.cnrs.fr/jigsaw">
          <stderr>
http://localhost:80/~bwali/Test2_13178161
12859-6806/std.err
          </stderr>
          <stdout>
http://localhost:80/~bwali/Test2_13178161
12859-6806/std.out
          </stdout>
          <simpleoutput1>
http://localhost:80/~bwali/Test2_13178161
12859-6806/ex2output1.nii
          </simpleoutput1>
          <simpleoutput2>
http://localhost:80/~bwali/Test2_13178161
12859-6806/ex2output2.nii
          </simpleoutput2>
</ns1:localResult>
```

(4)Result of the automatic generation of OWL-S grounding description based on the jGASW WSDL

```
<grounding:wsdlInputMessage
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://i3s.cnrs.fr/jigsaw#local
</grounding:wsdlInputMessage>
<grounding:wsdlInput>
<grounding:WsdlInputMessageMap>
<grounding:wsdlMessagePart
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-1.1.1/jigsaw?wsdl#simpleinput1
</grounding:wsdlMessagePart>
<grounding:owlsParameter rdf:resource="#Atomic_Process_Test2-
1.1.1_input1"/>
</grounding:WsdlInputMessageMap>
</grounding:wsdlInput>
<grounding:wsdlInput><grounding:WsdlInputMessageMap>
<grounding:wsdlMessagePart
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-1.1.1/jigsaw?wsdl#simpleinput2
</grounding:wsdlMessagePart><grounding:owlsParameter
rdf:resource="#Atomic_Process_Test2-1.1.1_input2"/>
</grounding:WsdlInputMessageMap>
</grounding:wsdlInput>
<grounding:wsdlDocument
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-1.1.1/jigsaw?wsdl
</grounding:wsdlDocument>
<grounding:owlsProcess rdf:resource="#Atomic_Process_Test2-
1.1.1"/>
<grounding:wsdlOutputMessage
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://i3s.cnrs.fr/jigsaw#localResponse
</grounding:wsdlOutputMessage>
<grounding:wsdlOutput><grounding:WsdlOutputMessageMap><grou
nding:owlsParameter rdf:resource="#Atomic_Process_Test2-
1.1.1_output1"/><grounding:wsdlMessagePart
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-.1.1/jigsaw?wsdl#localResult
</grounding:wsdlMessagePart>
</grounding:WsdlOutputMessageMap></grounding:wsdlOutput>
<grounding:wsdlOperation><grounding:WsdlOperationRef>
<grounding:operation
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-
1.1.1/jigsaw?wsdl#local</grounding:operation>
<grounding:portType
rdf:datatype="http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost:8080/Test2-
1.1.1/jigsaw?wsdl#jigsawPort</grounding:portType>
</grounding:WsdlOperationRef></grounding:wsdlOperation>
```

Figure 1: Automatic grounding of a jGASW service using the WSDL2OWLS API

This figure shows the main problem faced when we try to automatically generate the semantic description of a jGASW service; Columns 3 and 4 show how inputs are grounded individually whereas outputs are grounded as a single box “locaResult”. Therefore, OWL-S Process and Profile of this service contain one single output according to the obtained grounding. Why doesn’t OWL-S understand different outputs? Because they belong to a complex type “jigsawOutputTest2111” showed on column 2. Such output types generated automatically by the jGASW software can be even more complex than that. Actually, they can contain multiple nestings of complexType.

Motivation: First, we have shown that WSDL files are only partially understood by OWL-S API. Second, we need *control construct* to design and execute our image processing workflows. Third, OWL-S is a well defined language for web services composition that offers many functionalities that are not handled by other languages. Moreover it is submitted in the W3C , something important in our context of collaborative research in neuro-imaging. This choice of OWL-S is also consistent with our use of the OWL-Lite OntoNeuroLOG ontology as domain ontology. Indeed, OWL-S is an OWL-DL ontology and it is more expedient to use ontologies that are closer to each other in term of semantic capability and reasoning (i) both are based in OWL and this facilitate the use the same kind of reasoner (ii) if we use WSMO for example we should translate OntoNeuroLOG in WSML. Finally, OWL-S provides the suitable expressivity for representing web service semantics and fits nicely our application's requirements; in fact we cannot modify the structure of WSDL files, nor the XML description of inputs/outputs, since they are intrinsic to the jGASW middleware (otherwise invocation would not work properly).

Approach: First we extend the OWL-S Profile to be adapted to OntoNeuroLOG ontology. Second, we extend OWL-S Process to enable the description of jGASW services and finally, we add a software layer that implements reasoning services ensuring various consistency checks based on knowledge imbedded in the OntoNeuroLOG ontology.

III. Method

The OntoNeuroLOG ontology describes the different kinds of brain images in reference to the Dataset taxonomy and the functionality of services in reference to the Data processing taxonomy, each of these classes having its specific characteristics defined using DL axioms.

A. Extending OWL-S

OWL-S is a particular OWL ontology. It allows the semi-automatic composition of Web services. It is composed of three layers. The Service Profile allows the description, publication, and discovery of services. It is used by providers to publish their services and by users to specify their needs. The Service Model is used to compose services. It allows modeling services as processes. Three types of processes exist: atomic processes (AtomicProcess), simple (SimpleProcess) and composite (CompositeProcess). AtomicProcess represents the finest level of action that the service may perform. Composite Process are decomposable into other processes thus their concatenation can be specified using a set of control structures such as Sequence, Split, If-Then-Else, etc. A SimpleProcess is used to provide a view of an atomic process or a simplified representation of a composite process.

The Service Grounding describes how to access the service and provides the mapping between semantic inputs, outputs, message formats, and physical addresses. The purpose of this mapping is to enable the translation of semantic inputs generated by a service consumer into the appropriate WSDL messages for transmission to the service provider, and the translation of service output messages back into appropriate semantic descriptions (i.e., OWL descriptions) for interpretation by the service consumer.

1) Extending the OWL-S Process model

Our extension aims at decomposing the output grounded as a single box (localResult) with OWL-S (described in Figure 1) into its different elements (e.g., stderr, stdout, simpleoutput1, simpleoutput2). To overcome this problem, some classes and data/object properties are added to the OWL-S process model:

- a NlogParameter class to denote parameters that are embedded in a parameter of such a composite nature (e.g., stderr, stdout, simpleoutput1, simpleoutput2); it is defined as:

```
<owl:Class rdf:about="#NlogParameter">
  <rdfs:subClassOf rdf:resource="#Parameter"/>
  <owl:disjointWith rdf:resource="#Output"/>
</owl:Class>
```

- a nlogExpandsTo object property, associating a parameter of a composite nature to its essential elements according to the XSD schema of the service;

```
<owl:ObjectProperty rdf:about="#nlogExpandsTo">
  <rdfs:domain rdf:resource="#Output"/>
  <rdfs:range rdf:resource="#NlogParameter"/>
</owl:ObjectProperty>
```

- a hasID data property, denoting the markup within the string result after the service invocation

```
<owl:DatatypeProperty rdf:about="#hasID">
  <rdfs:domain rdf:resource="#NlogParameter"/>
  <rdfs:range rdf:resource="&xsd;anyURI"/>
</owl:DatatypeProperty>
```

- a hasLabel data property, denoting a non-functional property providing an informal description of the parameter

```
<owl:DatatypeProperty rdf:about="#hasLabel">
  <rdfs:domain rdf:resource="#NlogParameter"/>
  <rdfs:range rdf:resource="&xsd;anyURI"/>
</owl:DatatypeProperty>
```

Figure 2 provides an illustrative example of the use of the previous extensions: we present one extension of the output of jGASW service 2.

```
<process:Output rdf:about="&tool;#Atomic_Process_Test2_output1">
          <process:nlogExpandsTo><process:NlogParameter
rdf:ID="Atomic_Process_Test2_output1_simpleoutput1">
<process:hasLabel rdf:datatype=
"http://www.w3.org/2001/XMLSchema#anyURI">
          This is the extension of the first parameter simpleoutput1
</process:hasLabel> <process:hasID rdf:datatype=
"http://www.w3.org/2001/XMLSchema#anyURI">
          simpleoutput1
</process:hasID> <process:parameterType rdf:datatype=
"http://www.w3.org/2001/XMLSchema#anyURI">
http://localhost/dataset-owl-lite.owl#T1-weighted-MR-dataset
</process:parameterType>
</process:NlogParameter></process:nlogExpandsTo> </process:Output>
```

Figure 2: Enrich the output in its essential parameters

To detect the value of "simpleoutput1" argument first we select its ID using the hasID property and second we parse the string result (SOAP envelop) using the retrieved ID corresponding to the markup of "simpleoutput1". The data type of the NlogParameter is given by the data property *Process:parameterType.*

2) *Extending the OWL-S Profile model*

As described before, the OWL-S Profile gives information about the capabilities and the behavior of the service. We enrich it by adding a reference to the equivalent data processing class using refers-to, an object property that belongs to OntoNeuroLOG.

```
<owl:ObjectProperty rdf:about="&iec;refers-to">
    <rdfs:domain rdf:resource="#Profile"/>
    <rdfs:range rdf:resource="&data-processing-owl-lite;data-processing"/>
</owl:ObjectProperty>
```

3) *Web services composition*

As mentioned in the background section we cannot modify the WSDL otherwise invocation would no longer work, in consequence we could not extend the Grounding sub-ontology. The extension of the Process Model is enough to allow jGASW services composition. Once OWL-S Outputs have been related to corresponding *NlogParameters* according to the XSD Schema derived from jGASW processing, we were able to compose jGASW services. To this end, we introduced another object property, *links*, that binds any OWL-S Parameter to another (also suitable to NlogParameter since they are a subClassOf Parameter):

```
 <owl:ObjectProperty rdf:about="#links">
            <rdfs:domain rdf:resource="#Parameter"/>
            <rdfs:range rdf:resource="#Parameter"/>
</owl:ObjectProperty>
```

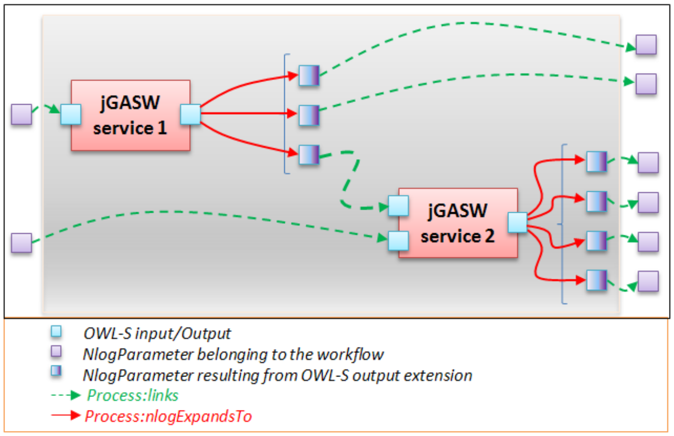

jGASW service 1
jGASW service 2
□ OWL-S input/Output
□ NlogParameter belonging to the workflow
□ NlogParameter resulting from OWL-S output extension
--→ Process:links
→ Process:nlogExpandsTo

Figure 3 : How to link NlogParameters with OWL-S parameters and workflow parameters in case of workflow composition

In this illustrative example, we compose two jGASW services: the first one service 1) has one input and one output. The output is composed of three outputs according to its XSD Schema, and the second (service 2) has 2 inputs and one output. The output is composed of four outputs according to its XSD schema. The profile of the service embedding the whole workflow has two inputs linked respectively to jGASW service 1 and jGASW service 2 and six outputs coming from both jGASW services. One internal parameter only is transmitted from service 1 to service 2.

B. *Some reasoning mechanisms*

1) *Compatibility check between dataset processing and the OWL-S Profile*

This service allows users to ensure that the definition of the profile is compatible with the data processing class selected by the user at annotation time.

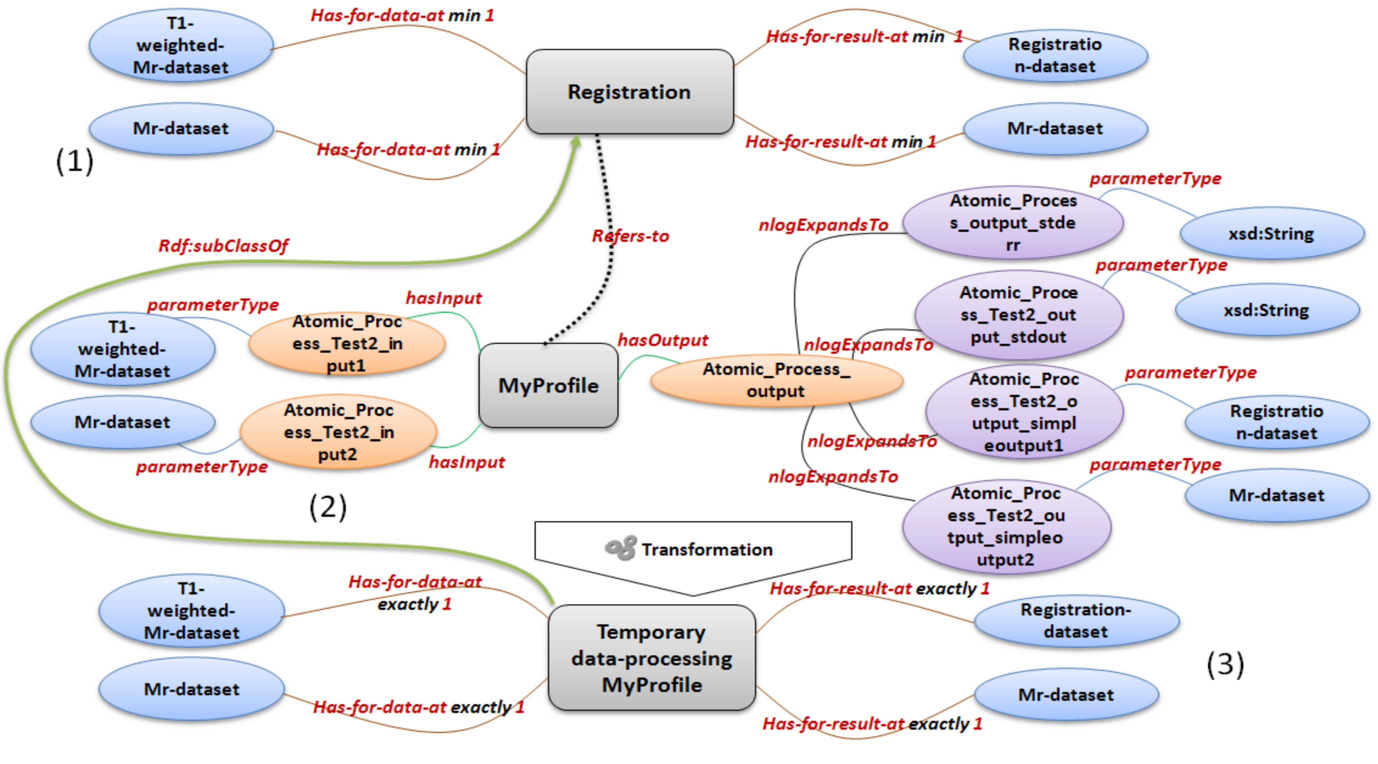


Figure 4: Transformation of Profile to data processing class

(1) Represents the description of Registration data processing, (2) represents the semantic description of the registration tool according to enriched OWL-S that should do registration if invoked and (3) shows the transformation of the profile into data processing.

The algorithm is the following: first we create a temporary class *tmp_Profile_data-processing* class relatively to the current operation, and then we translate relations between profile, inputs, and outputs into axioms and we add profile them to the *tmp_Profile_data-processing* class. Then, for every relation hasInput/hasOutput we count the number of inputs grouped by dataset class to determine the cardinality of the corresponding axiom; for example: *Process:hasInput* **i1** *Process:parameterType* Mr-dataset and *Process:hasInput* **i2** *Process:parameterType* Mr-dataset would lead to a cardinality of 2 concerning Mr-dataset (Mr-dataset denotes a magnetic resonance image dataset). The third step consists in selecting the appropriate object property for the construction of the axiom (e.g. *Process:hasInput* substituted by *has-for-data-at* and *Process:hasOutput* substituted by *has-for-result-at*. The result of the two first steps is: (*Process:hasInput* **i1** *Process:parameterType* **Mr-dataset** and *Process:hasInput* **i2** *Process:parameterType* **Mr-dataset**) ➔ (*has-for-data-at* **exactly 2** Mr-dataset). The third step consists of adding those axioms to the *tmp_Profile_data-processing* class.

The last step is to add the new tmp_Profile_data-processing class with axioms added above as subclass of the class referred by the Profile "MyProfile" and selected by the user, in our example (tmp_Profile_data-processing subclassOf Registration), and then, classify and check consistency. If the ontology is consistent then the annotation is considered valid. Semantically, the functionality of the tool is agreed, i.e., the has-for-data-at/has-for-result-at object properties are consistent with respective inputs/outputs specified in the corresponding data processing class in the OntoNeuroLOG ontology. Figure 4 show an illustrative example of the algorithm.

2) Compatibility check between outputs and inputs in a workflow

This service is applied when a user builds a new workflow. The processing aims at ensuring for every link between NlogParameter and Input that corresponding types are compatible. So we distinguish three cases:

- Identical data types: the output and the input have exactly the same type. Compatibility is validated and composition is accepted.
- Link to a more specific data type: the output is more general than the input of the next service, so non-compatibility.
- Link to a more general data type: the output is more specific than the input of the next service. The first service will always return results that are semantically compatible with the next service input. Compatibility is validated and composition is accepted.

N.B. workflow is valid if Parameters have the same Type or source is subsumed by target according to the dataset ontology.

3) Compatibility check between values and inputs at invocation time

This service is called when a web service is invoked. It checks whether the actual instances selected by the user (e.g. a Dataset) and assigned to the values actually meet the constraints specified in the semantic annotations of the service. In practice, the semantic service checks whether the class (or the type) of this instance is subsumed by the class type of the input.

IV. IMPLEMENTATION

The semantic annotation of jGASW services is generated automatically using the WSDL2OWLS API. Enrichment of semantic annotation is done using the OWL-S 1.2 specification and the OWLS API 3.0.

The semantic annotation of workflow services is generated using the OWL-S 1.2 specification and the OWLS API 3.0.

The consistency check between the profile and the data processing class is implemented using the OWL API, the OWL-S API 3.0 and the HermiT Reasoner.

The web services invocations use the OWL-S API but results and composition issues use the semantic search engine CORESE [21] together with the OWL-S API. CORESE is used to select the functional properties (extensions, linked parameters, identifiers …) of OWL-S outputs by querying the triple store containing the semantic annotations of the services. We add here an illustrative example of a workflow composed of two services (Figure 5). First, we prepare the SOAP envelope to invoke the first jGASW service: green markup shows the WSDL operation input (tns:local) and blue markup indicates the concrete input that the service will use.

```
<soapenv:Envelope><soapenv:Body>    <local    xmlns="http://i3s.
cnrs.fr/jigsaw"><simpleinput    xsi:type="xsd:string"    xmlns="http:
//i3s.cnrs.fr/jigsaw">http://localhost/test1.nii</simpleinput>
</local></soapenv:Body></soapenv:Envelope>
```

The next section shows the output of the service after invocation: green markup shows the output of the WSDL operation (tns:localResult). It wraps three blue markups that show three files generated by the service.

```
<ns1:localResult                    xmlns:ns1="http://i3s.cnrs.fr/jigsaw">
<stderr>http://localhost:80/~bwali/Test1_1321350928548-9787/std.err
</stderr>   <stdout>http://localhost:80/~bwali/Test1_1321350928548-
9787/std.out</stdout><simpleoutput>http://localhost:80/~bwali/Test1
_1321350928548-9787/testoutput.nii</simpleoutput>
</ns1:localResult>
```

The stderr and stdout are workflow outputs whereas simpleoutput should be transmitted to the second jGASW service. With CORESE we query the triple store to retrieve the nlogParameters to which the output (localResult) is extended. Then, for every nlogParameter retrieved, we find the ID (the markup to extract it from the localResult), its link to another parameter (parameter passing), and its data type. The parameter that we extract is simpleoutput. It should be transmitted to ex002_input2 second input of the second jGASW service.

Query: aims at identifying the different outputs that tns:localResult (corresponding semantically to *ex001_output1*) is expanded to:

```
PREFIX p1: <http://localhost/kb/Test1_2.owl#>
PREFIX p2: <http://localhost/Process.owl#>
Select   ?nlogParameter   ?link   ?id   ?type   where  {
p1: ex001_output1   p2:nlogExpandsTo   ?nlogParameter
?nlogParameter      p2:links           ?link
?nlogParameter      p2:hasID           ?id
?nlogParameter      p2:parameterType      ?type }
```

Query Results:

```
?nlogParameter  http://localhost/kb/extension-Test1_2.owl#ex001
_simpleoutput
?link  http://localhost/kb/Test1_2.owl#ex002_input2
?id  simpleoutput
?type http://localhost/dataset-owl-lite.owl#T1-weighted-MR-dataset
?nlogParameter  http://localhost/kb/extension-Test1_2.owl# ex001
_stdout
?link  http://localhost/kb/extension-Test1_2.owl#WF_stdout
?id  stdout
?type  http://www.w3.org/2001/XMLSchema#string
?nlogParameter  http://localhost/kb/extension-Test1_2.owl# ex001_
stderr
?link  http://localhost/kb/extension-Test1_2.owl#WF_stderr
?id  stderr
?type  http://www.w3.org/2001/XMLSchema#string
```

The output ex001_output1 is expanded to three nlogParameters as seen in the Figure 5 (ex001_stdout, ex001_stderr, ex001_simpleoutput) corresponding respectively to (stdout, stderr, simpleoutput) in the query results (?id fields). Those ID are the markups used in the localResult. The query results show that both ex001_stdout, ex001_stderr are linked to workflow outputs (WF_stdout1, WF_std_err1) as showed in the Figure 5. The query results show that the parameter ex001_simpleoutput is linked to the parameter ex002_input2.

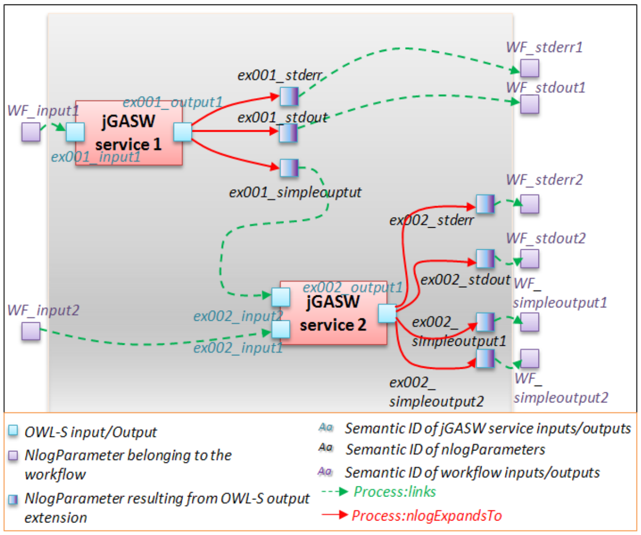


Figure 5: semantic annotation of workflow using the OWL-S Process layer and the extension described in this work

Thus it should be passed to the second jGASW service. To this end, the value of *ex001_simpleoutput* is extracted using the jGASW engine by giving the ID already selected by the query. The result is:**http://localhost:80/~bwali/Test1_1321350928548-9787/testoutput.nii.** A new SOAP envelope containing two inputs (as Figure 5 shows) is prepared to invoke the second jGASW service.

```
<soapenv:Envelope        <soapenv:Body>><local        xmlns="http://
i3s.cnrs.fr/jigsaw"><simpleinput1        xsi:type="       xsd:string"xmlns=
"http://i3s.cnrs.fr/jigsaw">http://localhost/test4.nii        </simpleinput1>
<simpleinput2    xsi:type="xsd:string"    xmlns="http://i3s.cnrs.fr/jigsaw">
http://localhost:80/~bwali/Test1_1321350928548-9787/testoutput.nii
</simpleinput2> </local></soapenv:Body></soapenv:Envelope>
```

The simpleinput1 is the file selected by the user for the workflow execution (corresponding semantic id is *WF_input2*). This parameter is passed to *ex002_input2*. The simpleinput2 gets the file extracted from localResult. i.e. result of the execution of first jGASW service invoked as described above. The other parameters (stderr, stdout, simpleoutput1, and simpleoutput2 of jGASW service2) are transmitted to the workflow outputs.

V. DISCUSSION

Several semantic languages and frameworks have been proposed based on W3C web service languages to support web service composition. However, web service composition is hampered by the heterogeneity of web services. Our work is an extension of OWL-S at the concrete service level to address the issue of jGASW web services composition.

We relied on OWL-S because it is a well defined Ontology [21] based on manifold earlier solutions and it is currently submitted in the W3C. It is also a semantic framework that provides more complete specifications than any other alternative solutions. It is represented in OWL which is a standardized language and exploits its reasoning capability [22]. Thus, it enables us to leverage our domain ontology in reasoning aiming at performing various consistency checks regarding the use of our services. OWL-S is a multi-layered language thus, it is easy to handle. In our contribution, extending the Profile layer and the Process layer leverages this characteristic. OWL-S differs from other specifications by providing conditions, effects, sequences and control constructs. We reused conditions and effects definitions to verify the consistency of service compositions and control construct specifying the behavioral aspect [23] of composed jGASW services. The OWL-S Service Grounding is conceived to be adapted for grounding any kind of service. Unfortunately, our WSDL files are really specific and cannot be grounded entirely. Getting service output as a unique box and as a string format actually hampers generating the grounding automatically and therefore the semantic description. Nevertheless, OWL-S is still the nearest solution and its adoption and extension allowed overcoming the problem.

WSDL-S [24] and SAWSDL [25] define how to add semantic annotations to WSDL specifications. In fact, they let WSDL components refer to semantic concepts via the **ModelReference** attribute, added to WSDL elements to assign one or more semantic concepts, via the **schemaMapping** property to map complex types and elements with a semantic model, via **Precondition** and **effect** for service discovery, via **serviceCategory** to help in case of service advertisement. In contrast to OWL-S they externalized domain application and let the reasoning mechanisms free. Grounding should be interpreted manually and service composition is not explicit. They do not deal with context of execution, behavior aspect and therefore, the reasoning aspect is really neglected, so we preferred use a more sophisticated and developed language for reasoning mechanisms.

Web service composition is still a complex task [1][26][27]. Numerous surveys on web service composition present an overview of methods that deal with web service composition. Based on a large background, Dustdar and Schreiner [27] discussed the need of web service composition and related issues. They outline the importance of the *context* in web service composition. The *context* should be formatted in some customized and personalized manner for relevant use by the next service. In our work we had to face the same requirements regarding the composition problem. The enrichment of OWL-S aims to format outputs in order to make them adequate for the next service that will be invoked. Enabling jGASW services composition is the added value of this enrichment and key factor of our work. It enabled us to add algorithms to check consistency

Rao and Su [1] investigated automated web service composition and propose an abstract framework for automatic service composition. They discuss abstract process model and business workflow involving the impact of *heterogeneity* of web services sources. We conclude that web service composition becomes more difficult if ever we deviate from the standard cases to specific cases. For example, automatic selection, matching, and composition work well while using standards. It is against hindered if we are out of standards. In our work, jGASW WSDL files are different from standard WSDL files. They differ by their XML schema, thus, they are heterogeneous compared to standard ones. This shows nevertheless, the dependency of semantics model on thin technical details and with the manner how to access services.

Without OWL-S the composition of jGASW service is not possible. In fact the form of SOAP envelop of the result not allows the chaining of web services. If we would like to compose jGASW services without semantics we should add interoperability within the jGASW engine. The first benefit from extension and use of OWL-S facilitate this task by enable the composition process. The second benefit from OWL-S is the multilayered structure that it has. In fact with the ServiceProfile it enables us to add semantic verification according to the neuro-imaging expectations. This is shown throw the implementation of the validation algorithm.

Casati et al. [28] uses the notion of process template to model composite services and composers need to browse the process library to search for process templates of interest [27]. Rao et al. [1] and Dustdar et al. [27] distinguish in workflow composition static and dynamic workflow generation, static defines the business tasks and dynamic linking the concrete e-services. Both help for monitoring e-services. OWL-S does not provide explicit support for monitoring and errors handling [29]. OWL-S service profile is just a service categorization and still lacks semantics. In our work we add some semantics to augment workflow monitoring. For example while users compose their workflows our consistency checking algorithm verifies that the service profile and the related data processing are consistent, which ensures that the chaining of the service can make sense from the point of view of processing.

Cardoso and ShethIn [30] try to overcome e-workflow composition problems by making services interoperable. They use a multidimensional approach based on ontology mediation. Medjahed et al. [26] address the interoperability issue by using composability rules. Currently, this task must be performed by a human who might use a search engine to find a service, and connect the service manually. However, a couple of verification algorithms were implemented within our application framework using OWL-S markup of services, and the necessary information from the OntoNeuroLOG ontology. At this stage, our work is still basic, the automatic discovery and mediation process are not well handled. Indeed, this process requires further development to overcome the heterogeneity of semantic web services using mediation. Especially WSMO, that uses the mechanism of mediation between semantic services coming from different heterogeneous frameworks. In our work the semi-automatic composition does not need mediation, however it needs a semantic validation through a reasoning aspect implementing verification of the consistency of the flows.

Gannod et al. [31] the authors present a generic approach to ground services with OWL-S. Users can ground automatically or manually the service to its description. Although it is a generic approach this kind of grounding does not meet our needs. In fact, it considers that every end point (WSDL or others) defines the outputs individually. However in our case the outputs are embedded in the unique box and are not explicit for the WSDL API and are understandable only by our jGASW Engine. Semantically, this editor considers that service grounding and service model are two distinct layers. In our work, we no longer keep those two layers separate, which is a limitation of our solution. In fact the process:hasID data property that was added is the unique way to access to the WSDL elements (input/output) as explained in the implementation section . We are required to do that because the way jGASW gets back the result obliges us to have a link to the parameter in the process specification. Otherwise, if we try to extend service grounding, invocation would no longer work.

Web services differ in form, technique or design point of view. Application wrappers provide outputs and inputs in different forms due to the functional requirements of the application domains. In this case, semantic solutions are not enough. The extension that we proposed is adequate for every kind of service so we augment the flexibility of web service development. Even the service has different technical details, the proposed idea, when reused in another context, is still valid and address both technical and semantic problems.

VI. Conclusion And Future Work

In this paper, we introduced a method to extend the OWL-S specification to cope with jGASW web services description. We succeeded to address the problem of semantic web services composition and to add some semantic validation and verification mechanisms. This solution addresses several issues concerning the web services composition in the neuro-imaging domain, but it is not sufficiently tested by NeuroLOG users. For us to assess its added value from an end user point of view, moreover, automatic composition still needed.

The next step of this work tends to, ensure and validate this work by adding serious test through the neuro-imaging framework and developing an algorithm for automatic selection and composition of jGASW web services and OWL-S workflows. We are trying to add reasoning capability over the description of data processing and the validation with profile algorithm.

ACKNOWLEDGMENT

The NeuroLOG project was funded by the French National Agency for Research (ANR-06-TLOG-024). The authors gratefully acknowledge all contributors, including our clinical colleagues providing the image data exploited in the NeuroLOG test-bed.

REFERENCES

[1] J. Rao, and X. Su, “A Survey of Automated Web Service Service Composition Methods,” In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition, SWSWPC’2004, LNCS, San Diego, USA, Springer-Verlag, July, 2004, pp. 43-54.

[2] A. Tsalgatidou, and T. Pilioura, “An Overview of Standards and Related Technology in Web Services,” journal of Distrib. Parallel Databases, Hingham, MA, USA, vol. 12, September, 2002, pp. 135-162.

[3] F. Curbera, R. Khalaf, N. Mukhi, S. Tai, and S. Weerawarana, “The next step in Web services,”journal of Commun. ACM, New York, NY, USA. vol. 46, October, 2003, pp. 29-34.

[4] http://uddi.xml.org/ (25-11-2011)

[5] M.P. Papazoglou, and W.V.D. Heuvel, “Service oriented architectures: approaches, technologies and research issues,” presented at VLDB J., 2007, pp. 389-415.

[6] T. Andrews et al.: "Business Process Execution Language for Web Services Version 1.1". Available at http://xml.coverpages.org/BPELv11-May052003Final.pdf, last accessed 20/01/2012.

[7] http://www.w3.org/TR/soap/ (22-11-2011)

[8] R. Chinnici, J. Moreau, A. Ryman, et al.,“Web Services Description Language (WSDL) Version 2.0,”Online: http://www.w3.org/TR/wsdl20/, June, 2007.

[9] E. Sirin, J. Hendler, and B. Parsia, “Semi automatic composition of web services using semantic descriptions,” in proceedings of the ICEIS-2003 Workshop on Web Services: Modeling, Architecture and Infrastructure, Angers, France, April, 2003.

[10] S. A. McIlraith, T. C. Son, and H. Zeng, “Semantic Web Services,” IEEE Intelligent Systems, March/April, 2001, pp. 46-53.

[11] http://www.daml.org/briefings/ (22-11-2011)

[12] http://www.daml.org/2001/03/daml+oil-index (20-12-2011)

[13] http://www.w3.org/TR/owl-features/ (22-11-2011)

[14] http://www.wsmo.org/ (01-02-2011)

[15] Martin, D., Paolucci, M.; Mcilraith, S.; Burstein, M.; Mcdermott, D.; Mcguinness, D.; Parsia, B.; Payne, T.; Sabou, M.; Solanki, M.; Srinivasan, N. & Sycara, K. Bringing Semantics to Web Services: The OWL-S Approach Springer, 2004, pp. 26-42.

[16] http://neurolog.i3s.unice.fr/neurolog (23-02-2012)

[17] F. Michel, A. Gaignard, F. Ahmad, C. Barillot et al., “Grid-wide neuro-imaging data federation in the context of the NeuroLOG project,”HealthGrid’10 (HG’10), IOS Press, 2010.

[18] J. Rojas Balderrama, J. Montagnat, and D. Lingrand, “jGASW: A Service-Oriented Framework Supporting High Throughput Computing and Non-functional Concerns,” IEEE International Conference on Web Services (ICWS 2010), IEEE Computer Society, Miami, FL, USA, July, 2010, pp. 5-10.

[19] L. Temal, M. Dojat, G. Kassel, and B. Gibaud, “Towards an Ontology for Sharing Medical Imagesand Regions of Interest in Neuro-imaging,” J. of Biom. Inf., vol.41, 2008, pp. 766-778.

[20] http://semwebcentral.org/projects/wsdl2owl-s/ (20-01-2012)

[21] O. Corby, R. Dieng-Kuntz, and C. Faron-Zucker, “Querying the Semantic Web with Corese Search Engine,” In: Proceedings of the 15th ECAI/PAIS, Valencia, Spain, 2004.

[22] Z. H. Yang, J. B. Zhang, and C. P. Low, "Towards Dynamic Integration of Collaborative Virtual Enterprise using Semantic Web Services," In Proceedings of the 4th International IEEE Conference on Industrial Informatics, Singapore, 2006.

[23] L. Guo, Y.Chen-Burger, and D.Robertson, “Mapping to business process model to semantic web service model,” In Proc. of the IEEE International Conference on Web Services, 2004, pp. 0-3.

[24] http://www.w3.org/Submission/WSDL-S/ (12-01-2012)

[25] http://www.w3.org/2002/ws/sawsdl/ (26/03/2002)

[26] B. Medjahed, A. Bouguettaya, and A.K. Elmagarmid, “Composing Web Services on the Semantic Web,” The VLDB J., vol. 12, no. 4, 2003, pp. 333-351.

[27] S. Dustdar and W. Schreiner, “A Survey on Web Services Composition,” Int'l J. Web and Grid Services, vol. 1, no. 1, 2005, pp. 1-30.

[28] F. Casati, S. Ilnicki, L. Jin, V. Krishnamoorthy, and M. C. Shan, “Adaptive and dynamic service composition in eFlow,” In Proc. of the CAiSE conference, Stockholm, June, 2000, pp. 13-31.

[29] R. Vaculín, and K. Sycara, “Monitoring execution of OWL-S web services,” In European SemanticWeb Conference, OWLS: Experiences and Directions Workshop, June, 2007, pp. 3-7.

[30] J. Cardoso, and A. Sheth, “Semantic e-Workflow Composition,” tech. report, Large Scale Distributed Information Systems Lab, Dept. of Computer Science, Univ. of Georgia, Athens, Ga., vol. 21, 2002, pp. 191-225.

[31] G. C. Gannod, R. J. Brodie and J. T. E Timm, “An Interactive Approach for Specifying OWL-S Groundings” *EDOC*, 2005, pp. 251-260.

Ontologies for Intelligent Provision of Logistics Services

Andreas Scheuermann
Information Systems 2
University of Hohenheim
70599 Stuttgart, Germany
andreas.scheuermann@uni-hohenheim.de

Julia Hoxha
Karlsruhe Institute of Technology (KIT), Karlsruhe
Service Research Institute (KSRI), Institute of Applied
Informatics and Formal Description Methods (AIFB)
76128 Karlsruhe, Germany
julia.hoxha@kit.edu

***Abstract*—The complex and volatile global economic environment challenges Supply Chain Management and increases the need for advanced Information Technology. To enable flexible and intelligent management of supply chains, we present an overall approach based upon a combination of Semantic Web Technologies and Service-oriented Computing. This work developes dedicated logistics ontologies for enabling intelligent provision of logistics services.**

Keywords-Logistics Ontology; Ontology Engineering; Semantic Technology; Service-oriented Computing; Supply Chain Management.

I. INTRODUCTION

The 21st century global markets become increasingly turbulent and volatile: life-cycles shorten due to the global economic environment and ever growing individual customer demands, while competitive forces put additional pressure on companies supply chains. Lengthy and slow-moving supply chains bounded by rigid organizational structures endanger companies' competiveness. In such a complex and dynamic environment, logistics excellence has become a powerful source of competitive advantage [1].

However, the relentless effort of companies to strive for competitive advantage has ballooned complexity of logistics decision-making and intensified the need to flexibly provide logistics capabilities. This development highlights the vital role of Information Technology (IT) and especially the need for flexible IT architectures and intelligent approaches for exploiting logistics knowledge [2][3].

From an IT viewpoint, current technologies such as Semantic Web Technologies (SWT) and Service-oriented Computing (SOC) bear considerable potential to enhance Supply Chain Management (SCM). SOC, as a paradigm for distributed, potentially cross-organizational software systems, aims at rapidly and easily providing applications by combining single services to enable flexible business processes. In this paradigm, a service is a loosely coupled, autonomous, and platform-independent computational entity, which encapsulates discrete functionality and can be accessed by using established web standards [4][5]. SWT allows for describing consensual knowledge of a specific domain of interest by means of formal semantics. This enables automated reasoning, information integration, semantic interoperability, and, thus, the application of intelligent approaches [6], e.g., for decision support or discovery and composition of logistics services. While each of these technologies has revolutionized IT, so far, the capabilities and advantages of combining both approaches have been rudimentarily exploited.

The objective of this paper is to extend the previous approach [7], which combines the paradigm of SOC and SWT, by applying them both on the logistics domain and SCM to enable intelligent provision of semantic logistics services. We propose a refined three-layered semantic approach consisting of a logistics semantic layer containing dedicated logistics ontologies, a logistics service description layer, and a logistics process description layer. From the perspective of ontology engineering, the contributions of this work are dedicated logistics ontologies that provide formal logistics knowledge to unambiguously describe logistics services and artefacts. In contrast to state-of-the-art logistics ontologies, the proposed ontologies (1) sufficiently capture the logistics domain not only restricted to specific logistics areas, and (2) incorporate a higher degree of formal semantics.

The rest of the paper is organized as follows: Section 2 reviews related work. Section 3 introduces the underlying approach. Section 4 develops dedicated logistics ontologies, which are evaluated in Section 5. Section 6 draws a conclusion and points to future work.

II. STATE-OF-THE-ART

A query to dedicated ontology search engines (e.g., http://swoogle.umbc.edu) reveals that the actual number of logistics ontologies is very small. The available ontologies can be assigned either to the domain of manufacturing or exclusively to specific areas of logistics (e.g., aircraft types, IATA codes, hazardous cargo). These ontologies merely provide taxonomies lacking formal axioms.

In scientific publications, the work of Wendt et al. [8] presents aspects of merging two domain-specific ontologies (production logistics and hospital logistics) to derive common logistics concepts for scheduling and facilitate efficient communication processes. The ontology itself is not published. Chandra and Tumanyan [9] apply an ontology to systematically record knowledge about organizational and problem-specific issues for SCM. They propose an information modelling framework to create a taxonomy of supply chain problems and operations to alleviate operational uncertainty. Madni et al. [10] introduce the IDEON ontology as a basis for designing, reinventing, managing, and controlling collaborative and distributed enterprises. IDEON integrates multiple perspectives, such as an enterprise context view or a process view. It is represented using the Unified Modelling Language. Lin et al. [11][12] develop a manufacturing system engineering (MSE) ontology to support an intelligent coordination tool within extended or virtual enterprises. The MSE ontology conforms to a taxonomy of different concepts: project, enterprise, process, extended enterprise, resource, and strategy. These ontologies are modelled by means of software engineering techniques, conform to simple taxonomies, and merely address specific logistics aspects.

Another group applies ontology languages upon existing logistics models ("ontologizing"). Fayez et al. [13] propose an OWL representation of the SCOR model for supply chain simulation. The ontologies should capture the distributed knowledge being required to integrate several supply chain views in order to support the construction of simulation models. Leukel and Kirn [14] develop a logistics ontology based on the SCOR model to capture core concepts of inter-organizational logistics. The proposed ontology facilitates the description of activities in logistics and provides relations and attributes. While these publications provide richer ontologies, they are still limited to few abstract concepts.

The last group aims at extending well-grounded ontologies. Haugen and McCarthy [15] propose an extension of the REA Ontology to support internet supply chain collaboration. Pawlaszczyk et al. [16] introduce an ontology based on the Enterprise Ontology to describe the domain of mass customization for optimizing inter-organizational and distributed cooperation. The ontology introduced by Soares et al. [17] focuses on production planning and control in a virtual enterprise environment to improve human communication and to support the specification of system requirements. The ontology is founded on the meta-ontology of the Enterprise Ontology, whereas the concepts are defined by natural language and object models. Ye et al. [18] propose a supply chain ontology to enable semantic integration between heterogeneous supply chain information systems. The supply chain setting is a web-based or virtual enterprise with no specific industry focus. The ontology is implemented in OWL and based on the Enterprise Ontology. This group of ontologies concentrates on supply chains rather than on a larger scope of the logistics domain. However, all listed ontologies lack rich formal semantics, remain rather abstract, merely focus on special logistics aspects, and neglect the service-oriented nature of logistics.

III. RESEARCH DESIGN

The overall approach underpinning our work is to combine SWT with SOC and apply them on the logistics domain. The ultimate goal is to enable the intelligent and flexible provision of logistics services in supply chains and customized logistics applications (Fig. 1).

SWT [6] comprises formal knowledge representation and reasoning capabilities. Thereby, ontologies provide appropriate means to formally structure and explicate knowledge about the logistics domain in order to enable semantic interoperability, information integration, and reasoning.

SOC [19], particularly Web Services, conform to modular, loosely-coupled software components that are accessible by established web standards and, thus, facilitate the provision of services in distributed and heterogeneous environments. Web Services allow for flexible business process management, which when combined with formal semantics provide capabilities for (semi-)automated service discovery, ranking, composition.

The combination of SWT and SOC lay the basis for semantic logistics services, which encapsulate discrete logistics functionality (e.g., transport), consume logistics resources, and whose quality is measurable by logistics performance indicators. Semantic logistics services consist of modular, reusable, and loosely-coupled logistics components for flexibly managing complex logistics processes.

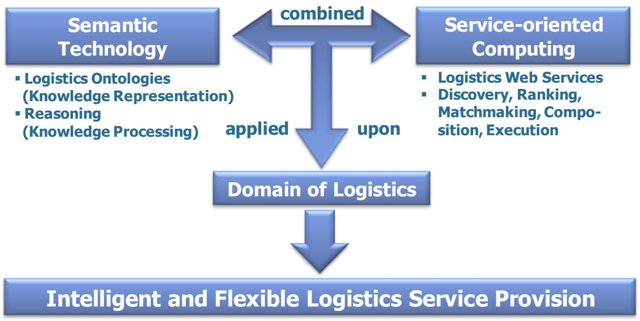

Semantic Technology
combined
Service-oriented Computing
▪ Logistics Ontologies (Knowledge Representation)
▪ Reasoning (Knowledge Processing)
applied
upon
▪ Logistics Web Services
▪ Discovery, Ranking, Matchmaking, Composition, Execution
Domain of Logistics
Intelligent and Flexible Logistics Service Provision

Figure 1. Combining SWT and SOC for SCM

Based on this approach, we introduce a three-layered model for engineering Semantic Logistics Services (Fig. 2).

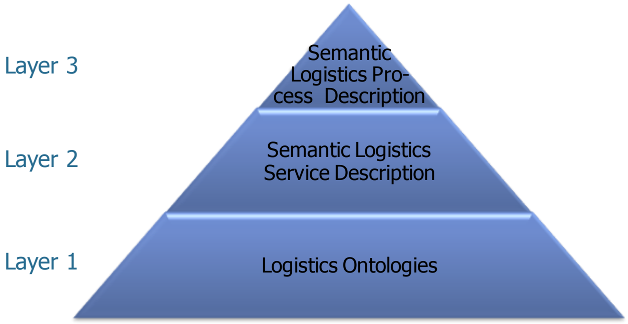

Layer 3
Semantic Logistics Process Description
Layer 2
Semantic Logistics Service Description
Layer 1
Logistics Ontologies

Figure 2. Three-Layered Model for Semantic Logistics Services

Layer 1: Logistics Ontologies build the foundation for defining formal semantics of consensual logistics knowledge. We provide a modular ontological setup, which allows for easy reuse, adaption, and refinement.

Layer 2: Semantic Logistics Service Descriptions are used for the representation of atomic logistics services. We exploit

OWL-S Service Profile [20] for the description of service features and utilize the logistics ontologies of Layer 1 for semantic annotation.

Layer 3: Atomic logistics services are composed into complex logistics processes, which are semantically described by SuprimePDL [21] process description language. Layer 3 consists of such Semantic Logistics Process Descriptions.

IV. ENGINEERING LOGISTICS ONTOLOGIES

We focus on Layer 1 to propose dedicated logistics ontologies for semantically describing logistics services.

A. *Scenario: The Case of Fourth- Party Logistics*

The increasing pressure on companies to rapidly and flexibly provide superior logistics capabilities has led to logistics outsourcing, which involves both operational logistics services (e.g., transport) and management capabilities (e.g., planning). As an optimal solution, the concept of fourth-party logistics (4PL) has emerged. A 4PL is challenged to provide logistics expertise and IT to integrate and manage – plan, implement, and control – logistics services in complex and dynamic supply chains without possessing own logistics assets [22]. In particular, a 4PL could significantly benefit from exploiting both SWT and SOC:

First, the use of the logistics ontologies and the OWL-S Service Profile allows 4PL and other logistics service providers (LSP) to unambiguously characterize their offered services. Accurately fulfilling customer service levels directly affects a retailer's or manufacturer's competitive ability. Exploiting formal semantics not only fosters semantic interoperability and (semi-) automatic data integration between heterogeneous logistics information systems along the supply chain but also speeds up communication and makes it more flexible and error-resistant. Further, logistics ontologies allow for automated reasoning to expose implicit knowledge. Customer requirements on logistics service provision dynamically change, thus, reasoning enables to check usability of services beyond explicitly stated characteristics.

Second, applying semantic logistics services allows for advanced functionality with regard to logistics service discovery, ranking, matchmaking, and composition. A 4PL faces both a huge amount of end-customers with individual requirements concerning logistics service provision and, moreover, a variety of LSPs for satisfying these requirements. Hereby, logistics service discovery and ranking accelerate matching demand with supply of logistics services. As supply chain complexity and dynamicity constantly increase, service composition enables a 4PL to rapidly and flexibly integrate logistics services, and, thus, to configure and manage highly complex supply chains.

B. *Logistics Domain Capture*

The ontology engineering approach is based on a dedicated methodology [24]. A main characteristic of this methodology is the formulation of informal competency questions (CQ) to determine the scope and purpose of the ontology. These questions incorporate the terminology and functional requirements of the ontology to be developed. In our case, the development of CQ is based on two main factors. First, the CQ relate to logistics theory in terms of logistics literature and standards (e.g., UN/CEFACT, SCOR). Second, domain experts – developers of logistics software, logistics experts and managers – participated in workshops to contribute to the development of the CQ. Table I depicts some general CQ, which are further extended and substantiated.

TABLE I. COMPETENCY QUESTIONS

CQ1:	What actors participate in the provision of logistics services?
CQ2:	What roles can logistics actors play?
CQ3:	What types of logistics services are offered by LSP?
CQ4:	Which functional and nonfunctional parameters characterize logistics services?
CQ5:	Which metrics characterize logistics service performance?
CQ6:	Which logistics units and goods flow through supply chains?
CQ7:	Which resources are needed for logistics services provision?

C. *Logistics Ontology Modeling*

The objective of the logistics ontologies is to capture and structure overall knowledge of the logistics domain to annotate logistics services. The logistics ontologies from [7] are further extended and modularized to facilitate reusability, extensibility, and maintainability. To encode the logistics ontologies we apply OWL 2 DL [25].

Figure 3 zooms into Layer 1 and provides an overview of the modularly organized logistics ontologies, their concepts, and relations.

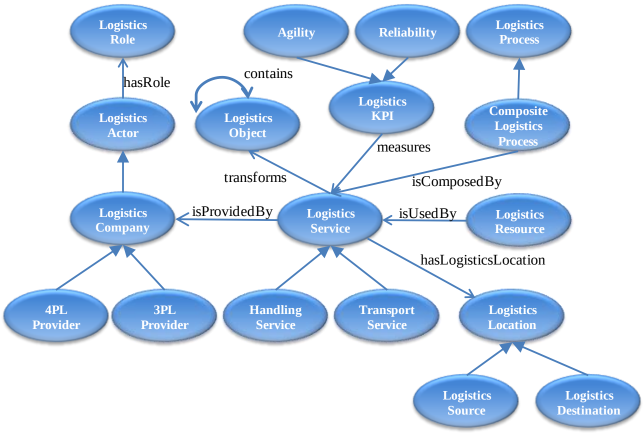


Figure 3. Logistics Ontologies

The main component of the logistics ontologies is the Logistics Service Ontology. Its fundamental concept is LogisticsService that encapsulates spatial, temporal, and quantitative transformations according to the logistics basic functions: transport, handling, warehouse, and complementing value-added services. Logistics companies provide logistics services that respectively represent a temporally and factually logic finite sequence of states to completely transform a logistics object.

$$\mathit{LogisticsService} \equiv \mathit{TransportService}\ \sqcup$$

$$\text{HandlingService} \sqcup \text{WarehouseService}$$
$$\text{LogisticsService} \sqsubseteq \forall\text{isProvidedBy.LogisticsCompany} \sqcap \exists\text{transforms.LogisticsObjects}$$

The Logistics Process Ontology is especially important with respect to Layer 3 in order to handle dynamic aspects of real-world logistics complexity. Its key concept is Logistics-Process, which comprises atomic or composite logistics processes. Thereby, a composite logistics process conforms to a composition of at least two logistics services. For instance, a transport process contains various logistics services (e.g., transport, handling) being composed to realize the pre-, main, and on-carriage of a supply chain.

$$\text{LogisticsProcess} \equiv \text{AtomicLogisticsProcess} \sqcup \text{CompositeLogisticsProcess}$$

Subject to logistics transformations are Logistics Objects. According to the logistics unit-load concept, logistics objects are a combination of economic goods with charge carriers. For instance, microprocessors (goods) are packaged in boxes (charge carrier), which are in turn consolidated on pallets and/or containers. The unit-load concept aims at a smooth flow of logistics objects from source to destination.

$$\text{LogisticsObject} \sqsubseteq \text{ChargeCarrier} \sqcup \text{Good} \sqcup \exists\text{isTransformedBy.LogisticsService}$$
$$\text{ChargeCarrier} \equiv \text{Container} \sqcup \text{Box} \sqcup \text{Pallet}$$

A supply chain consists of actors, which are defined in the Logistics Actor Ontology. Logistics actors are either individuals or corporative actors, i.e., legal entities created for business ventures. To model logistics actors, we consider the way they participate in supply chains. For instance, we distinguish between logistics companies dealing with the provision of logistics services and manufacturers aiming at the production of goods.

$$\text{LogisticsActor} \equiv \text{LogisticsCompany} \sqcup \text{Authority} \sqcup \text{ManufacturingCompany} \sqcup \text{TradingCompany}$$
$$\text{LogisticsCompany} \sqsubseteq \text{2PLProvider} \sqcup \text{3PLProvider} \sqcup \text{4PLProvider}$$

Further, we introduce the Logistics Role Ontology to depict the capabilities and responsibilities of logistics actors in a supply chain. The concept LogisticsRole is inevitable because a logistics actor could have different roles at a certain point in time or over a particular time period, e.g., a manufacturer may simultaneously act as a requester of logistics services or supplier of goods to another manufacturer.

$$\text{LogisticsRole} \sqsubseteq \text{ServiceProvider} \sqcup \text{ServiceRequester} \sqcup \text{Supplier} \sqcup \text{OEM}$$

Whereas logistics services conform to the process organization (dynamic component) of a supply chain a Logistics Location Ontology represents the structural organization of supply chains. Therefore, we combine general location concepts such as country, zip code, city, and street with specific logistics concepts, e.g., source, and destination. For instance, the availability and performance of a transport service is strongly determined by the logistics object, its source, and its (final) destination (e.g., location of a manufacturer).

$$\text{LogisticsLocation} \sqsubseteq \text{LogisticsSource} \sqcup \text{LogisticsDestination}$$

Providing logistics services requires Logistics Resources. Logistics resources are factors of production, which are used during logistics service provision. They can be conceptualized according to the different types of logistics services. For instance, providing a road transport service for heavy goods requires a truck being capable to transport goods with a weight greater than 50 tons.

$$\text{LogisticsResource} \sqsubseteq \text{TransportationResource} \sqcup \text{WarehouseResource} \sqcup \exists\text{isUsedBy.LogisticsService}$$

Moreover, the availability and capability of logistics resources affects logistics service performance. Key performance indicators (KPI) measure logistics service performance. These KPI are modelled in the Logistics KPI Ontology and represent business key figures that assess the degree of logistics service performance. For this purpose, we reuse parts of SCOR [26] to provide a detailed classification of logistics performance indicators.

$$\text{LogisticsKPI} \sqsubseteq \text{Agility} \sqcup \text{Costs} \sqcup \text{Reliability}$$

Beyond, we reuse and extend existing ontologies either specific to the domain of logistics or ontologies with a more general character. For instance, to model hazardous goods we reuse the hazardous goods ontology. Additionally, we reuse an ontology containing airport codes to unambiguously define airports as locations. To represent units of measurement we reuse and extend the units of measurement ontology (MUO).

V. EVALUATION

Ontologies are complex engineering artefacts. Their evaluation is crucial to fully put their potential into practice, to make them reusable, and maintainable. For instance, incorrect and low quality ontologies might not be readable due to vocabulary and/or syntax errors, and, in the case of incorrect semantics, they might not be usable by reasoning engines. The evaluation includes logistics vocabulary, semantic interoperability, and information integration.

A. Vocabulary of the Ontologies

Evaluation of the vocabulary of the logistics ontologies is based on the CQ and performed as a classroom experiment that simulates expert evaluation. The class room experiment comprised a review group of twenty persons. The group was composed of ten university students with a strong background in logistics and ten logistics practitioners. The review group was given the task of comparing and assessing the concepts/relations of the logistics ontologies with the keyword index of selected logistics textbooks and standards. To capture the results of the class room experiment, we used match strength. To achieve operationalization, we applied an ordinal scale with: 1st quartile below 25% (poor), 2nd quartile between 26% and 50% (satisfactory), 3rd quartile between 51% and 75% (adequate), and 4th quartile between 76% and 100% (good). For instance, the match strength 'adequate' displays that between 51% and 75% of the concepts and relations contained in the respective logistics ontology appear in the key word index of logistics textbooks. Homonym and synonym conflicts were dissolved.

TABLE II. OVERVIEW OF VOCABULARY EVALUATION

Logistics Ontologies	Logistics textbook [27]	Logistics textbook [28]	Logistics standard [29]	Logistics standard [30]
Logistics Process	adequate	adequate	adequate	poor
Logistics Service	good	good	good	adequate
Logistics Object	good	good	good	satis-factory
Logistics Actor	good	good	satis-factory	poor
Logistics Role	satis-factory	satis-factory	satis-factory	poor
Logistics Location	good	good	adequate	poor
Logistics Resource	satis-factory	good	adequate	satis-factory
Logistics KPI	satis-factory	adequate	adequate	poor

The experiment shows empirical evidence (Tab. II) indicating that across all developed logistics ontologies the match strength is good in 34% (11 out of 32), adequate in 25% (8 out of 32), satisfactory in 25% (8 out of 32), and poor in 16% (5 out of 32). Occurring deviations could be explained by (1) the transfer of concepts originating in SOC to the logistics domain, which in particular holds for the concepts of the Logistics Process Ontology, (2) terminological heterogeneities existing in logistics literature, and (3) the human factor when performing such experiments. In particular, the results in column 5 are influenced by the specificity of the corresponding logistics standard.

B. *Semantic Interoperability and Information Integration*

The logistics ontologies constitute a consensual terminological basis for semantic annotation of logistics services. Since logistics implies communications beyond organizational borders at a global range, there exists an urgent need of integrating heterogeneous and distributed data sources, in particular, to achieve semantic interoperability. The logistics ontologies should support the mediation between heterogeneous data sources and enable unambiguous service annotations. To illustrate the problem related to semantic interoperability, we review common terms and definitions used in conventional service descriptions of three real-world LSP and show respectively how they relate to concepts in the logistics ontologies.

TABLE III. SEMANTIC INTEROPERABILITY

Ontology Concept	LSP_1	LSP_2	LSP_3
Transport Service	Complete load	Full load	Truck load
Logistics Objects	General cargo	Parceled goods	Piece goods
2PLProvider	Shipper	Haulage contractor	Forwarder
Transportation Resource	Truck	Lorry	Commercial vehicle
Logistics KPI	Lead time	Period of supply	Time of delivery

Based on the content of Table III, we present an example originating from real-world data to demonstrate the evaluation of semantic interoperability in the proposed logistics ontologies. Thereto, we pose queries against the logistics ontologies formulated in SPARQL [31]. Originally designed as a query language for graph patterns in Resource Description Format (RDF), SPARQL is practically also used to encode queries against OWL knowledge bases, interpreting the basic graph-matching capabilities by using the semantics of the ontology language.

Example: A manufacturing company requests a 4PL to provide logistics capabilities, which correspond to the capabilities of the logistics company type 'shipper'. To find all names of logistics companies that provide such capabilities, a query is formulated as follows:

```
PREFIX lo: http://www.interloggrid.org/
LogisticsOntology.owl#
 SELECT ?logisticsCompany ?logisticsCompanyName
 FROM <http://www.interloggrid.org/
 LogisticsOntology.owl#>
 WHERE {
   ?shipper rdfs:subClassOf ?2PLProvider.
   ?2PLProvider rdfs:subClassOf ?logistics
      Company.
   ?logisticsCompany lo:hasFirm lo:logistics
      CompanyName.}
```

The output of the query, comprises names of Logistics Company instances of LSP_1, LSP_2, and LSP_3. This is due to the fact that we established equivalence among the classes Shipper, Forwarder, and HaulageContractor, being all subclasses of 2PLProvider.

VI. CONCLUSION AND FUTURE WORK

The paper proposed dedicated logistics ontologies to semantically annotate logistics services. The overall approach combines SWT and SOC, applying them on the logistics domain for flexible and intelligent SCM. We propose a three-layered model for engineering semantic logistics services. This model includes elements to describe both declarative as well procedural aspects. The main focus is on the foundation of the three-layered model constituted by dedicated logistics ontologies. The logistics ontologies are modularly organized and capture the overall concepts of the logistics domain. The application of these ontologies fosters semantic annotation of logistics services and facilitates semantic interoperability, information integration, and reasoning capabilities allowing for intelligent applications. The evaluation comprises aspects of the logistics ontology vocabulary, semantic interoperability, and information integration. This work provides logistics ontologies for advanced and more flexible provision of logistics services to enhance dynamic configuration of complex supply chains. This would be of particular benefit to all participants in cooperative logistics scenarios assuming that the logistics ontologies are used to describe (annotate) logistics services.

Future work includes an extensive documentation and the full integration, as well as application of the logistics ontologies in a logistics platform prototype.

REFERENCES

[1] M. Christopher, “The Agile Supply Chain: Competing in Volatile Markets,” *Industrial Marketing Management*, vol. 29, Jan., 2000, pp. 37-44, doi:10.1016/S0019-8501(99)00110-8.

[2] P. Helo and B. Szekely, “Logistics information systems. An analysis of software solutions for supply chain co-ordination,” *Industrial Management & Data Systems*, vol. 105, Jan. 2005, pp. 5-18, doi:10.1108/02635570510575153.

[3] J. Leukel, A. Jacob, P. Karaenke, S. Kirn, and A. Klein, “Individualization of Goods and Services: Towards a Logistics Knowledge Infrastructure for Agile Supply Chains,” Proceedings AAAI Spring Symposium (AAAI), Mar. 2011, pp. 36-49.

[4] B. Iyer, J. Freedman, M. Gaynor, and G. Wyner, “Web services: enabling dynamic business networks,” *Communications of the Assocation of Information Systems*, vol. 11, 2003, pp. 525-554, http://aisel.aisnet.org/cais/vol11/iss1/30.

[5] M. Papazoglou and W.-J. van den Heuvel, “Service oriented architectures: approaches, technologies and research issue,” *The VLDB Journal*, vol. 16, Mar. 2007, pp. 389–415, doi:10.1007/s00778-007-0044-3.

[6] R. Studer, R.V. Benjamins, and D. Fensel, “Knowledge engineering: Principles and methods,” *Data and Knowledge Engineering*, vol. 25, March 1998, pp. 161–197, doi:10.1016/S0169-023X(97)00056-6.

[7] J. Hoxha, A. Scheuermann, and S. Bloehdorn, “An Approach to Formal and Semantic Representation of Logistics Services, “ Proceedings of the ECAI’10 Workshop on Artificial Intelligence and Logistics (AILOG10), Aug. 2010, pp. 73-78.

[8] O. Wendt, T. Stockheim, S. Grolik, and M. Schwind, “Distributed Ontology Management Prospects and Pitfalls on Our Way Towards a Web of Ontologies”, Dagstuhl Workshop, 2002.

[9] C. Chandra and A. Tumanyan, “Organization and problem ontology for supply chain information support system,” *Data & Knowledge Engineering*, vol. 61, May 2007, pp. 263–280, doi:10.1016/j.datak.2006.06.005.

[10] A.M. Madni, W. Lin, and C.C. Madni, “IDEON TM: an extensible ontology for designing, integrating and managing collaborative distributed enterprises,” *Systems Engineering*, vol. 4, Feb. 2001, pp. 35–48, doi: 10.1002/1520-6858(2001)4:1<35::AID-SYS4>3.0.CO;2-F.

[11] H.K. Lin, J.A. Harding, and M. Shahbaz, “A Manufacturing system engineering ontology for semantic interoperability across extended project teams,” *International Journal of Production Research*, vol. 42, 2004, pp. 5099–5118, doi: 10.1080/00207540412331281999.

[12] H.K Lin and J.A. Harding, “A manufacturing system engineering ontology model on the se-mantic web for inter-enterprise collaboration,” *Computers in Industry*, vol. 58, June 2007, pp. 428–437, doi: 10.1016/j.compind.2006.09.015.

[13] F. Fayez, L. Rabelo, and M. Mollaghasemi, “Ontologies for supply chain simulation modeling,” Proceedings of the 37th Conference on Winter simulation (WSC’05), Dec. 2005, pp. 2364–2370.

[14] J. Leukel and S. Kirn, “A Supply Chain Management Approach to Logistics Ontologies in Information Systems,” Proceedings of the 11th International Conference on Business Information Systems (BIS08), Springer LNBIP, May 2008, pp. 95-105.

[15] R. Haugen and W.E. McCarthy, “REA, a Semantic model for Internet Supply Chain Collaboration,” ACM Conference on Object-Oriented Programming, Systems, Languages, and Applications, Business Objects and Component Design and Implementation Workshop VI: Enterprise Application Integration, Oct. 2000, http://jeffsutherland.org/oopsla2000/mccarthy/mccarthy.htm, <retrieved: March, 2012>.

[16] D. Pawlaszczyk, A.J. Dietrich, I.J. Timm, S. Otto, and S. Kirn, “Ontologies Supporting Cooperation in Mass Customization - A Pragmatic Approach,“. International Conference on Mass Customization and Personalization – Theory and Practice in Cebtral Europe, Apr. 2004, pp. 1-19.

[17] A. L. Soares, A.L. Azevedo, and J.P. De Sousa, “Distributed planning and control systems for the virtual enterprise: organizational requirements and development life-cycle,” *Journal of Intelligent Manufacturing*, vol. 11, June 2000, pp. 253–270, doi: 10.1023/A:1008967209167.

[18] Y. Ye, D. Yang, Z. Jiang, and L. Tong, “An Ontology-based Architecture for Implementing Semantic Integration of Supply Chain Management,” *International Journal of Computer Integrated Manufacturing*, vol. 21, Jan. 2008, pp. 1–18, doi: 10.1080/09511920601182225.

[19] M.N. Huhns and M.P. Singh, “Service-oriented computing: key concepts and principles,” *IEEE Internet Computing*, vol. 9, Jan./Feb. 2005, pp. 75–81, doi. 10.1109/MIC.2005.21.

[20] OWL-S: Semantic Markup for Web Services, http://www.w3.org/Submission/OWL-S/, <retrieved: March, 2012>.

[21] J. Hoxha and S. Agarwal, “Semi-automatic Acquisition of Semantic Descriptions of Processes in the Web,” Proceedings of The International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'10), IEEE, September, 2010, pp. 256-263, doi: 10.1109/WI-IAT.2010.271.

[22] M.A. Razzaque and C.C. Sheng, “Outsourcing of logistics functions: a literature survey,” *International Journal of Physical Distribution & Logistics*, vol. 28, Apr. 2008, pp. 89-107, doi . 10.1108/09600039810221667.

[23] A. Win, “The value a 4PL provider can contribute to an organisation,” *International Journal of Physical Distribution & Logistics*, vol. 38, 2008, pp. 674-684, doi: 10.1108/09600030810925962.

[24] M. Uschold and M. Grüninger, “Ontologies: Principles, Methods, and Applications,” *Knowledge Engineering Review*, vol. 11, Feb. 1996, pp. 93–155, doi. 10.1109/MIS.2002.999223.

[25] OWL 2 Web Ontology Language (OWL2), http://www.w3.org/TR/owl2-profiles/, <retrieved: March, 2012>.

[26] Supply-Chain Council. 2010. Supply Chain Operations Reference Model (SCOR®) Vers. 10.0, http://www.supply-chain.org, <retrieved: March, 2012>.

[27] H.C. Pfohl, Logistiksysteme. Betriebswirtschaftliche Grundlagen. Heidelberg et al., Springer, 2010.

[28] T. Gudehus, Logistik. Grundlagen, Strategien, Anwendungen. Berlin, Springer, 2005.

[29] Deutsches Institut für Normung (DIN): Transportdienstleistungen – Logistik- Glossar. DIN EN 14943:2006-03, 2006.

[30] Deutsches Institut für Normung (DIN): Transportkette-Grundbegriffe. DIN 30781 Teil 1, Teil 2, Beiblatt Teil 1, 1989.

[31] E. Prud'hommeaux and A. Seaborne, “SPARQL Query Language for RDF,” W3C Recommendation 15 January 2008, http://www.w3.org/TR/rdf-sparql-query/, <retrieved: March, 2012>.

Enhancing the Provability in Digital Archives by Using a Verifiable Metadata Analysis Web Service

Jan Potthoff, Sebastian Rieger
Steinbuch Centre for Computing
Karlsruhe Institute of Technology
Karlsruhe, Germany
[name.surname]@kit.edu

Paul Christopher Johannes
provet
University Kassel
Kassel, Germany
paul.johannes@uni-kassel.de

***Abstract*—In a variety of research areas the requirements for good scientific practice demand a proper long-term archiving of the data produced and handled throughout scientific processes. While the documentation was traditionally written in paper-based laboratory notebooks, the increasing amount of digital data and corresponding metadata has led to the development of electronic laboratory notebooks (ELN). To ensure the integrity and authenticity of the data special mechanisms have to be established while being stored in the ELN and digital archives for several decades. As different research areas use individual scientific tools in their processes, this paper describes a generic data verification system to enhance the provability of data and the corresponding metadata. The system was implemented using a Web service that uses multiple ingress, verification and egress modules. By using the Web service presented in this paper, the provability of scientific data in digital archives is profoundly enhanced. By automatically evaluating the probative force of the data and metadata and adding the system's digital signature as a result, the provability can be ensured by a separate third party that is trusted on the side of the ELN operator as well as the scientific community and jurisdiction.**

Keywords- ELN; evidence; digital archive; digital signature; provenance

I. Introduction

The amount of data that is generated and processed in scientific processes has been continuously growing [1]. Major drivers behind this trend are large scientific experiments, e.g., the LHC at CERN, which produces an immense volume of large-scale data. On the other hand more and more scientific institutions and funding organizations have issued guidelines for safeguarding good scientific practice that recommend a proper long-term archiving of the scientific data that led to published results. Besides protecting the bit stream, these practices also require a certain amount of compliance regarding the scientific process to ensure the long-term comprehensibility of the data. In this paper we introduce a solution that is able to enhance the provability of data being ingested in a digital archive by evaluating the metadata and checking the consistency. Results of the evaluation are logged and a model to quantify the verifiability and provability of the data (regarding the integrity and authenticity within the scientific process) is being presented in this paper.

To allow the integration with different scientific processes, tools and especially electronic laboratory notebooks (ELN) the solution was developed as a modular Web service. Different digital archives can also be connected to the Web service using specific modules. State of the art digital signatures are used to verifiably sign the evaluation results and to protect the provability in the connected archives. Besides the automatic verification and evaluation of metadata, where appropriate and possible, the system also checks digital signatures and timestamps that were generated by the scientist or archivist before being sent to our system.

In Section II, we give an overview on data in scientific processes, from the generation and processing up to publication and archival forming a scientific data lifecycle. We also give references to related work and related projects that focus on scientific long-term archiving, scientific provenance and provability of scientific records. In Section III, we describe the model that we developed to measure the probative force of scientific data and measures that we use to protect the authenticity and integrity of data while being stored in digital archives. Modules and the applied mechanisms that measure the probative force are described in detail in Section IV. The result from the data verification performed by these modules is quantified, classified and signed using the techniques described in Section V. In Section VI, we conclude our findings and discuss their impact on the scientific process as a link to future research.

II. Electronic Documentation In The Research Process

Documenting research digitally in ELN is not only a fad or de rigueur but state of the art practice in many fields of research [2]. Slowly ELN will replace the traditional lab journal made out of paper [3]. Their implementation into the research process presents challenges to scientist, research facilities, universities and technical staff [4]. This holds especially true for its seamless implementation into the research process and scientific data lifecycle.

A. Scientific Data Lifecycle

An assessment of the research process in general is difficult since it is highly individual and differs from scientific branch to branch and researcher to researcher. The quantity of different ELN available on the market [5] shows how distinct the methods of different scientific branches are. Still the scientific process flow may be roughly divided into five phases. At the beginning the experiments are planned on the basis of theoretical considerations and their set parameters (design). On this basis experiments are performed and then evaluated (implementation and processing). In the course this, effects are discovered and results are documented and thus serve oneself and others, possibly by subsequent publication (analysis and publication). Finally, the gathered data and results have to be archived so that they can be used for later reuse or for presentation (archiving). In relation to the origin, use, storage and reuse of the processed data we speak of the scientific data lifecycle [6].

B. Related Work

The law concerning record keeping in scientific research processes is spread among many statues and is diverse. It changes from jurisdiction to jurisdiction. Specialized work into the provability of scientific records is rare. Related work usually refers to specific scientific fields. Charrow [7], for example, covers the complex laws and generally accepted procedure that relate to biomedical research in the USA. In this paper, we will discuss the legal provability of ELN in the most universal and abstract way possible.

Improving the secure long-term archival of digitally signed documents is still a challenge [8]. Solutions not only depend on the available technology but also on the law of the applicable jurisdiction [9]. To record all past versions of scientific data, versions of the database should be preserved in a continuous fashion [10]. We will expand on this research and apply its findings to stored scientific research data.

The provenance (also referred to as the audit trail, lineage, and pedigree) of electronic scientific data should contain information about the process and data used to derive it. It is documentation that is key to preserving the data, to determine its quality and authorship and to reproduce as well as validate the results [11]. These are all important parts of the scientific process and scientific metadata and can be circumstantial evidence to the authenticity and integrity of electronic research data. There are strong arguments to preserve metadata for legal evidence as a regular practice [12]. We seek to create a viable solution for (automatically) creating, archiving and utilizing metadata that is generated in the scientific data lifecycle.

The challenge is to record uniform and usable provenance metadata that meets regulatory needs while minimizing the modification burden on the scientist and the performance overhead on scientist and system [13]. Minimizing setup and maintenance costs by automating the database design, data load, and data transformation tasks helps to seamlessly integrate an ELN into the research process [14]. Current ELN offer little or no direct support for "scientist-oriented" queries of provenance information [15]. We address this with our automated weighting and classification model.

III. Designing a System to Improve the Probative Force of Scientific Data

Any system for the electronic documentation of the research process and its data should be designed with the goal to ensure the legal provability of its content. As opposed to the scientific provability of theories by means of experiments or empiric studies, which design and scope always depends on the subject matter and the branch of science and relies on peer review, the need for legal provability stems from the concern of the scientist to prove his results not only to other scientists but also to other people and institutions, e.g., in a court of law. Reasons for this are manifold [6]. For example, a scientist could be accused of scientific deception [16] or scientific fraud [17]. If the data he presented is questioned, he will want to show, that the data was not invented, alternated, falsified or parts of it suppressed, in order to exonerate his credibility and exculpate himself from any wrongdoing or even criminal liability.

A. Digital Signatures and Timestamps

To authenticate an author of electronic research data and evidence the integrity of the record, certain kinds of digital signatures can be used. These digital signatures (i.e., advanced electronic signatures) are data in electronic form, which are associated with other electronic data and serve as a method of authentication. In order to have probative value, they must be uniquely linked to the signatory, be capable of identifying him and be linked to the data to which it relates in such a manner that any subsequent change in the data is detectable. Many countries have adopted rules to the legality and evidentiary value of digital signatures (see [24] for a comprehensive but not complete list; see also [18]). Within the EU, for example, the rules have been harmonized by Directive 1999/93/EC of the European Parliament and of the Council of 13 December 1999 on a Community framework for electronic signatures. The rules therein had to be adopted by all EU Member States. They ensure that mutual legal recognition of qualified certificates and electronic signatures from third countries is applied if certain reliability conditions are met.

The Directive defines different classes of electronic signatures. The main provision of the Directive states that an advanced electronic signature based on a qualified certificate created by a secure-signature-creation device satisfies the legal requirements of a signature in relation to data in electronic form in the same manner as a handwritten signature satisfies those requirements in relation to paper-based data (for convenience this type of signature is usually called a "qualified signature"). It is also admissible as evidence in legal proceedings.

Qualified signatures are recognized by procedural law in many EU-Member states and, due to their origin from a known, trusted party, which is monitored by a respective EU-Member state, of high probative value in a court of law. Therefore the integrity of an electronic document signed with a verifiable qualified signature can be proven as well as its authenticity. In the context of research data, a scientist can evidence that his electronic records originate from him and have not been altered since the last time he has signed them with his qualified signature. But in addition to that, a scientist will want to show that his data has not been altered since a definite point in time. This could, e.g., be the date of record, the day of the experiment, and the date of archival.

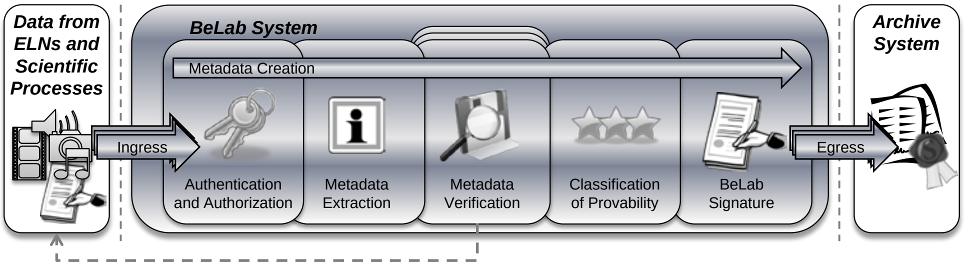


Figure 1. Modules of the BeLab Web Service

Later modification and falsification of electronic records can be ruled out by using trustworthy digital timestamps. According to the RFC 3161 standard, a trusted timestamp is a timestamp issued by a trusted third party (TTP) acting as a Time Stamping Authority (TSA). It is used to prove the existence of certain electronic data before a certain point without the possibility that the owner can backdate the timestamps. The newer ANSI X9.95 standard for trusted timestamps expands on RFC 3161 by adding data-level security requirements that can ensure data integrity against a reliable time source that is provable to any third party. Both standards can be used to create trustworthy timestamps that cannot be altered without detection and to sustain an evidentiary trail of authenticity. This holds especially true, if the TSA uses legally recognized digital signatures for his timestamps, like qualified signatures.

B. Evidentiary Value and Classification

A trail of authenticity, like a complete chain of evidence, is in the context of scientific research data a means to show that no scientific deception or fraud has been committed. To do so, the scientist wants to be able to prove that no data has been altered or falsified since recording it and no data has been suppressed. It is therefore advisable to ensure data integrity and authenticity as early as possible in the research process. The later the archival of research data, the greater is the possibility that the data has been tampered with before any kind of security measure like a digital signature or timestamp has been applied. Both, early application of digital signature and automation of the signing process are evidence for the integrity of the data [19].

But still: A general assessment of the evidentiary value of an electronic document is difficult to near impossible. The evidentiary value (i.e., probative force) always depends on the individual case: Who bears the burden of proof, i.e., which party needs to prove an assertion of fact, and whether the offered evidence is at all suitable to prove the fact [20]? In order to prepare for the eventuality of a legal dispute, the user of electronic records should assess their probative value. It is crucial to know how the authenticity and integrity of electronic documents can be proven and how this can be sustained [18]. Therefore it makes sense to give the user a comprehensible system for evaluating the authenticity, integrity and security of the dataset at hand. Such a system should classify the evidentiary value of data based on the data formats, the metadata collected and the kind of digital signatures used. Through the classification the scientist receives an indication of the degree of evidentiary value and potential risks of long-term archiving and preserving his research data or ELN.

C. Data and Metadata Model

As described in Section II, several different tools and ELN are typically used to manage data in scientific processes. Even in single scientific processes of a specific research area we found multiple ELN, and ordinary applications, e.g., Microsoft Office, being used together with custom software solutions, i.e., for data analysis. To support these data sources we implemented a Web service (called BeLab) that offers a generic input interface. Figure 1 depicts the modules and data processing to analyze and preserve the probative force of scientific data. By using HTTPS the confidentiality of the data transfer is protected.

After the authentication and authorization of the data transfer originating from the ELN (based on the user of the ELN or the ELN as a whole), metadata is extracted from the received data and stored in a container that is used during the subsequent verification and classification of the probative force. We chose an existing implementation that offers an extendible (XML-based) metadata container encoding & transmission mets standard. The ingress modules build up a TAR file of the data to be archived. By using a common standard the ELN can use a unified way to describe metadata information for the archived data.

The result of the execution of the verification modules, being stored in the model, is quantified in the classification module, which is explained in Section V. To protect the classification of the probative force the system signs the classification result using a digital signature. Finally the BeLab system offers a generic output interface, which supports different egress modules to digital archive systems that preserve the bit stream and the long-term interpretability of the data.

D. Archive systems

Multiple archive systems can be connected to the BeLab system by implementing specific egress modules. The implementation of the BeLab system also allows the combination of multiple archives, i.e., to ensure redundancy and fault tolerance across the archives. Typically these measures against bit stream errors or manipulation of the stored data are already addressed within the long-term archive system. An example for a standard that implements such a mechanism that uses digital signatures and digitally signed timestamps is described in RFC 4998 as long-term archive and notary service. Archive systems that are receiving data from the BeLab system can verify the digital signature of the BeLab system in their ingress interface or during a cyclic refresh of the long-term archived data. The egress modules of the BeLab system serialize the data being stored in the BeLab model as described in the previous section and transmit the data to the archive.

IV. Automatic Weighting Modules

As described in Section III.C, the probative force of the data being submitted to the BeLab system is processed and stored using a unified model inside the BeLab system. Ingress modules supply the data to be archived and associated metadata using a TAR file that includes a mets XML file containing the metadata. To enhance the quality of the automatic verification of the data and metadata carried out by the BeLab system, an additional metadata extraction is performed before the execution of the verification modules. This also allows the selection of the proper verification modules that support the specific data or metadata format or content. Furthermore by comparing the results of the metadata extraction to the supplied metadata an initial consistency check of the submitted data is performed.

A. *Integrity by Checksums*

The metadata contains unsigned hash value of each file in the archive file, as described in Section III. To ensure the integrity of the submitted data the checksum module calculates the hash value of each file again and compares it to the value that was received from the ingress module. In order to use the same algorithm to calculate the hash value the algorithm name is specified in the mets container.

B. *Integrity and Authenticity by Digital Signatures*

The checksum module as described in the previous section does not suffice to ensure the integrity and authenticity of electronic documents. As described in Section III.A, digital signatures, i.e., based on a X.509 certificate, are needed to ensure the authenticity of the document. Some computer programs, e.g., Microsoft Office, OpenOffice and Acrobat, offer the opportunity to sign the corresponding document. In these examples the signature is integrated in the file format. Other programs, e.g., Cryptonit, offer the possibility to save the signature in a separate file. In this way it is possible to sign electronic documents which do not offer the integration of signatures. Before archiving signed data, the electronic signature and corresponding certificates should be verified. Therefore the BeLab system implements specific modules which are designed to support different signature standards, i.e., PKCS#7 and XML-DSig.

1) *Embedded Digital Signatures*

The digital camera from the company Kappa optronics GmbH [25] which can be used to collect data samples in scientific processes, offers the opportunity of electronically signed images. Therefore, the image will be signed before the data will be transferred to the computer that the camera is attached to. Using the software included with the camera different file formats can be selected, e.g., jpg, gif or bmp. The signed images have a data format that was developed especially by Kappa optronics GmbH. The verification module that makes it possible to check the signature has to understand this specific format.

Hence a specific (Kappa) module was implemented for the BeLab system. The file format is based on three parts which contain the image data with header information and custom content, the signature and the public key to decrypt the signature. In the first step the Kappa module calculates the hash value based on defined data range. After the signature decryption both hash values will be compared to verify the signature.

In contrast, the integration of signatures in PDF is based on PKCS#7. So, a corresponding weighting module can use frameworks which are already developed such as Bouncy Castle Crypto API [23]. With this framework it is possible to verify digital signatures based on PKCS#7 which can be used in the PDF weighting module. This module also allows an automatic verification of the integrity of the data, validity of embedded certificates and optionally contained signed timestamp [21].

Additionally a module was implemented that allows users to verify XML-DSig signatures [21]. This kind of signature is used, for example, in the OpenOffice file format. To verify the signature the Bouncy Castle Crypto API was used. The OpenOffice file format is based on a zip archive file. Hence the archive needs to be unpacked to verify the signature. The OpenOffice module unpacks the archive with the given structure and verifies the signature. Even though the verification of signatures based on PKCS or XML standard can be handled in a uniform manner individual modules are needed.

2) *External Electronic Signatures*

External digital signatures are typically based on a uniform standard (PKCS#7). Therefore a single module was implemented to check for each file whether a corresponding signature file has been supplied. The signature file has to match the file name and must be placed in the same folder as the associated file [21]. If an external signature has been found the module starts the verification, as described in Section IV.B.1).

C. *Sequence Detection and Workflow Verification*

Digital cameras produce images that are usually named on the basis of a sequential number. These numbers can be used to automatically verify the completeness of a received series of images. One or more missing files can be interpreted in two ways: The owner might have forgotten to submit all files or he might have been trying to alter the series of images [21]. For whatever reason the files are missing, the proprietor of the ELN should be informed about it, as this might have an impact on the probative force of the data. Therefore the corresponding weighting module should take the sequence detection into account [19].

The implemented module works in two steps: First it checks whether or not a series exists. Therefore all numbers in the provided filename will be removed. After that the remaining part of the filename will be compared with the other results from previous filenames. If there are more than x identical designations a new series of files, e.g., images, is found. The value x can be set individual by the user. If a series of images is found the module checks the completeness in a second step. First, the lowest and highest number of the series has to be determined. Next the series can be analyzed whether or not an image is missing. All steps of the sequence detection will be noted in a log [21].

V. Classification Modules Based on Logging Mechanism

Not all implemented verification and weighting modules are suitable for each file format. The right modules must be selected based on the corresponding file type. Therefore a specific module, that is able to determine the data format, was implemented [21]. As described, there are dependencies between different modules, meaning that some modules need the result of another module and should be executed in the right order. The module for determining the file format should be executed first.

For each module a value (index) can be defined which indicates the order of modules. In addition, there are other conditions for selecting a module, e.g., dealing with file archives or separate files. These conditions can also be defined in the weighting module definition.

A. The Verification Result as Log

Due to the modular approach of the BeLab system, the result of a weighting module should be flexible [22]. Therefore, the result model of the verification was implemented as a log which can include different results [21]. Each entry consists of three values: key, content and type of content. The key element defines the validation being executed, the content element includes the result of this validation, and the type element declares the data type of the result. For each module a set of keys is defined. For example, the module for embedded signatures, as described in Section IV, defines the keys: “signature found”, “signature verified” and “signer”. In this case the type of “signature found” and “signature verified” is Boolean and the type of “signer” is String.

Each weighting module can produces any number of log entries. More than one module can analyze one file. The result is a list of logs which is stored as a tuple of filename and result. The complete verification and logging process is shown in Figure 2.

B. Log Analysis to Determine the Probative Force

The coordination of the verification process is performed by a validation controller. First, the controller chooses the right weighting modules for the evaluation of the whole archive, e.g., the sequence detection module, as presented in Section IV. Appropriate modules are chosen by the corresponding file format after that. When the validation process is finished, the coordination of the classification process is performed by a separate controller. This controller chooses the classification modules the same way as the validation controller does it with the weighting modules. This ensures that each result of the verification is taken into account in the classification process. The goal is to map the results of the verification regarding the degree of probative force, the suitability for long-term preservation and the degree of secure (i.e., consistent) data generation.

The probative force is primarily determined by the use of digital signatures and digital timestamps. For example, if the embedded signature module, as described in Section IV, has found a signature, the corresponding classification module would check its validity based on the result of the embedded signature module. If it is valid the module checks the specific kind of signature in a second step. The result is a value from B1 standing for “no signature” to B6 the highest qualification based on the qualified electronic signature, as described in Section III [22].

The suitability for long-term preservation is based on the file format. Because there is no unambiguous approach to define the qualification there are some roles which can be used. For example, the complexity of the file format, the usage of open standards in an index for the suitability of a data format. In this case the result of the classification is L1 (inappropriate), L2 (appropriate), L3 (recommended) and can be understood as a recommendation, meaning that the BeLab system does not require specific file formats to support individual formats used in scientific processes [22]. However, the data container described in Section III must be used.

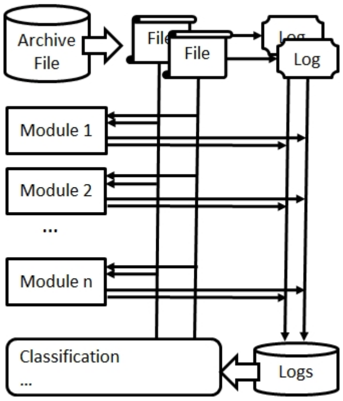


Figure 2. Verification process according to [21]

The degree of the secure data generation can be associated with the completeness of data [21], as shown in Section IV.D, or with the knowledge about the weighting modules used in the analysis phase. The completeness of data, for example, a series of digital pictures, indicates the integrity of this data and the scientific process. A stronger indication are electronic signatures which were used in the data generation phase [19]. For these signatures a corresponding weighting module, as shown in Section IV.B and classification module can be implemented. If the signature is valid, a secure data generation can be assumed. The BeLab system distinguishes the values S1 (insecure data generation) and S2 (secure data generation).

In fact, more than one module classifies each file resulting in different weighting modules being involved. Hence a mechanism to merge the classification results was needed. Initially, the final classification result is set to undefined (u). So, the classification for one file and the three categories, as described above, starts with a tuple <u, u, u>. If a module classifies a file, the result is another tuple, for example <u, B6, u>, meaning that the degree of probative force is set to the highest level of security and the other categories are undefined. These two results are merged by comparing each value of the three categories separately. If one of the values is undefined the defined value is preserved. In case both values are set, the lower value is taken. This way, a manipulation to a higher classification by a user is prevented.

All results of the weighting modules, the result of the classification process and eventually detected errors during the manufacturing process, are documented in a copy of the metadata. Finally, the metadata file and hence all included data (whether it was already signed or not) is signed by the BeLab system to ensure the integrity and traceability of the BeLab process. In addition, if multiple digital signatures for different documents were used, the verification is centralized. To do so, first the checksums have to be verified by the checksum module, as described in Section IV. Subsequently the signature of the metadata has to be verified, to ensure the integrity of submitted data.

VI. Conclusion and Future Work

By using ELN for the documentation of the scientific process, all data can be administrated centrally. Measures for the integrity and authenticity for paper-based laboratory notebooks cannot be transferred to the electronic documentation easily. By using the BeLab system, it is possible to ensure the integrity and authenticity of electronic data before the archive process. Using the developed weighting modules the needed data analysis can be executed automatically and without disturbing the scientist during his work. Additionally, the scientist gets useful information from the BeLab system during the archive process, i.e., about the suitability of the used file format. As the results from the evaluation of the BeLab system are stored in the attached digital archive, they can be used subsequently to prove the authenticity and integrity of the data. In Section IV.B.1), the example of a digital camera being used in scientific processes was given. One of the next requirements will be to support more measuring instruments and the validation of their data, e.g., using existing Web services.

References

[1] J. Gray, D.T. Liu, M.N. Santisteban, A. Szalay, D.J. DeWitt, and G. Heber, “Scientific data management in the coming decade,“ ACM SIGMOD Record, vol. 34, Dec. 2005, pp. 34-41, doi:10.1145/1107499.1107503.

[2] M. Kihlén, “Electronic lab notebooks – do they work in reality?,” DDT, vol. 10, Sep. 2005, pp. 1205-1207.

[3] M. Kihlén and M. Waligorski, “Electronic lab notebooks – a crossroad is passed,” DDT, vol. 8, Nov. 2003, pp. 1007-1009.

[4] D.J. Drake, “ELN implementation challenges,” DDT, vol. 12, Aug. 2007, pp. 647-649.

[5] M. Rubacha, A.K. Rattan, and S.C. Hosselet, “A Review of Electronic Laboratory Notebooks available in the market today,”, JALA, vol. 16, Feb.2011, pp. 90-98, doi:10.1016/j.jala.2009.01.002.

[6] S. Hackel, P.C. Johannes, M. Madiesh, J. Potthoff, and S. Rieger, “Scientific Data Lifecycle – Beweiswerterhaltung und Technologien,” Proc. 12. Deutscher IT-Sicherheitskongress (BSI-IT-SEC 2011), SecuMedia, 2011, pp. 403-418.

[7] R.P. Charrow, Law in the Laboratory. Chicago, IL: UCh. Press, 2010.

[8] C. Troncoso, D. De Cock, and B. Preneel, “Improving secure long-term archival of digitally signed documents,” Proc. 4th ACM international workshop on Storage security and survivability (StorageSS 08), ACM, 2008, pp. 27-36, doi:10.1145/1456469.1456476.

[9] W. Zimmer, T. Langkabel, and C. Hentrich, "ArchiSafe: Legally Compliant Electronic Storage," IT Professional, vol. 10, Jul.-Aug. 2008, pp. 26-33, doi:10.1109/MITP.2008.82.

[10] P. Buneman, S. Khanna, K. Tajima, and W. Tan, „Archiving scientific data,“ ACM TODS, vol. 29, Mar. 2004, pp. 2-42, doi:10.1145/974750.974752.

[11] S.B. Davidson and J. Freire, “Provenance and scientific workflows: challenges and opportunities,” Proc. ACM SIGMOD international conference on Management of data (SIGMOD 08). New York, NY:ACM, 2008, pp. 1345-1350, doi:10.1145/1376616.1376772.

[12] W.L. Wescott II, „The Increasing Importance of Metadata in Electronic Discovery,“, Rich.J.o.L.T-, vol. 14, Article 10, http://law.richmond.edu/jolt/v14i3/article10.pdf <retrieved: March 12, 2012>.

[13] B. Howe, K. Tanna, P. Turner, and D. Maier, “Emergent Semantics: Towards Self-Organizing Scientific Metadata,“ in Semantics of a Networked World, M. Bouzeghoub, C. Goble, V. Kashyap and S. Spaccapietra, Eds. Berlin, DE: Springer 2004, pp. 177-198, doi:10.1007/978-3-540-30145-5_11.

[14] S. Bowers, T. McPhillips, B. Ludäscher, S. Cohen, and S.B. Davidson, “A Model for User-Oriented Data Provenance in Pipelined Scientific Workflows,” in Provenance and Annotation of Data, L. Moreau and Ian Foster, Eds. Chicago, IL: Springer, 2006, pp. 133-147, doi:10.1007/11890850_15.

[15] Y.L. Simmhan, B. Plale, and D. Gannon, “A Framework for Collecting Provenance in Data-Centric Scientific Workflows,” in Proc. International Conference on Web Services (ICWS 06), IEEE Press, Sep. 2006, pp. 427-436, doi:10.1109/ICWS.2006.5.

[16] L. Grayson, Scientific Deception: an overview and guide to the literature of misconduct and fraud in scientific research. London, UK: British Library, 1990.

[17] H. Ottemann, Wissenschaftsbetrug und Strafrecht. Hamburg, DE: Dr. Kovač 2006.

[18] S. Mason (Ed.), International Electronic Evidence, London, UK: BIICL 2008.

[19] J. Potthoff, S. Rieger, P.C. Johannes, and M. Madiesh, “Elektronisch signierende Endgeräte im Forschungsprozess,” Proc. D-A-CH Security 2011, syssec, 2011, pp. 44-55.

[20] E. Schneider, Die Klage im Zivilprozess. Cologne, DE: Otto Schmidt 2007, pp. 477ff.

[21] F. Ellmer, “Automatische Metadatenanalyse zur beweiswerterhaltenden Langzeitarchivierung im Forschungsprozess,” Karlsruhe Institute of Technology, bachelor thesis.

[22] M. Madiesh, P.C. Johannes, and J. Potthoff, “Beweissichere elektronische Labor-, Patienten- und Fallakten,” Proc. perspegtive 2011, in press.

[23] Bouncy Castle, http://www.bouncycastle.org <retrieved: March 12, 2012>.

[24] eSignatureLegalWiki.org contributors, “Main Page,” eSignatureLegalWiki.org, http://www.esignaturelegalwiki.org <retrieved: March 12, 2012>

[25] Kappa optronics GmbH, http://www.kappa.de/ <retrieved: March 12, 2012>.

Towards Automatic Non-Deterministic Web Service Composition

George Markou
Dept of Applied Informatics
University of Macedonia
Thessaloniki, Greece
e-mail: gmarkou@uom.gr

Ioannis Refanidis
Dept of Applied Informatics
University of Macedonia
Thessaloniki, Greece
e-mail: yrefanid@uom.gr

***Abstract*— This paper describes our ongoing work towards the implementation of an online Web Service Composition system, based on the most prevalent Web Service standards and utilizing other open source projects as sub-elements. The system will treat non-determinism as an inherent characteristic of the problem and will tackle it by exploiting AI planning technology, specifically contingency planning. The paper also presents three detailed use case scenarios to evaluate the system's capabilities. The final system will be the first online application of its kind able to support various stages of Web Service Composition.**

Keywords-Web Service; NuPDDL; non-determinism; OWL-S; composition

I. Introduction

As Web Services (WSs) exist and operate in an ever-changing and expanding environment, it is difficult to expect from a human user, or even an expert, to manually or semi-automatically complete the goal of a Web Service Composition (WSC) process. The number of WSs is growing continuously and, as such, the WSs discovery phase becomes more difficult. Web Services can change interfaces or even part of their usage multiple times throughout their lifespan; even if they remain static, there is always the possibility that their execution is not successful. A WSC process should automatically detect and respond to such changes in a way that a human will probably not be able to.

This paper presents our ongoing work towards a WSC system that will exhibit the following functionalities:

- Advertisement of a new WS in a registry, as well as online editing and retrieval of the WSs already stored.
- Translation between the language used to describe the semantic WSs taking part in the composition and an Artificial Intelligence (AI) planning one, OWL-S and NuPDDL [1], respectively.
- Generation of a composition process model, based on contingency planning and OWL-S' control constructs.
- Evaluation of the WSC process, based on quantitative criteria (e.g., the number of WSs considered for the composition, the transformation time of the WSC domain to a planning one, or the total planning time) and pre-defined use case scenarios.

To our knowledge, no other open source web-based WSC system exists. Our system's implementation is based on existing freely available components so as to insure maximum conformance to the current WS standards and facilitate its quantitative evaluation and comparison to other WSC systems. So far, the advertisement functionality has already been integrated in our online application; currently, we are working on the implementation of a manual WSC module, as well as the complete mapping between the elements of OWL-S and NuPDDL.

The remainder of the paper is organized as follows: Section II reviews related work, while Section III presents the approach that is being followed to implement the various WSC functionalities of the proposed system. Section IV focuses on its evaluation; more specifically, it presents three scenarios that can be used as test cases for our system. Finally, Section V concludes the paper and poses directions for future work.

II. Related Work

Several approaches that convert the original WSC problem to a planning one have been proposed; perhaps the most notable is [2], in which the available WSs' OWL-S process models are translated to a SHOP2 domain, and the WSC problem to a compatible HTN planning problem by describing it as a composite process that can be decomposed to simpler ones (with simple processes being atomic WSs). Then, the planner generates a plan that can be converted back to an OWL-S process and executed by OWL-S API.

SHOP2 plans for tasks in the same order in which they will be executed, which allows it to be aware of the current state of the world at each step. In that way it can gain significant reasoning power in regard to its precondition-evaluation mechanism and reduce the complexity of the planning process. On the other hand, the approach cannot cope with non-determinism in the WSC problem and is planner-dependent, limiting it in comparison to others that translate the WSC problem to a PDDL-compliant one.

An approach very similar to [2] is presented in [3], which couples SHOP2 with an OWL-DL reasoner so as to tackle common problems of HTN methods, such as the inability to associate preferences with possible decompositions. Although the results presented are promising and the planner's performance is adequate, the approach is based on the initial provision of a template, which cannot always be known a priori. Moreover, the authors extend OWL-S with a new process type, thus hindering the use of existing test sets or tools. Finally, the use of a specific planner and internal representation excludes the use of alternative PDDL-compliant planners that could possibly tackle the problem more efficiently if a standardized presentation was adopted.

An AI planning methodology is also followed in [4], which treats the application of a WS as a belief update operation. Moreover, it identifies two special cases of WSC that are more tractable and allow for a compilation into planning under uncertainty and the subsequent use of an already existing conformant planner (Conformant-FF). Again, though, despite the use of planning techniques and a planner that takes as input PDDL-like problem descriptions, no standardized WS description or planning language is used.

PDDL and OWL-S are, respectively, the de facto planning language and the most widely used semantic description language. Moreover, the latter has been heavily influenced from planning languages, such as PDDL, and for that reason, a (perhaps partial) mapping from OWL-S to PDDL is relatively natural and intuitive. As such, there are several attempts that utilize them together in WSC problems.

OWLS-Xplan [5] incorporates a conversion tool that translates OWL-S descriptions to corresponding PDDL 2.1 ones; this translation, though, does not output a standard PDDL file, but a modified version of it in XML, which, according to the authors, simplifies parsing, reading, and communicating PDDL descriptions using SOAP. Then, Xplan, a hybrid planner that combines guided local search with graph planning and a simple form of HTN decomposition, is called to solve the planning problem. A similar approach is adopted in [6], which, however, can use two alternative external PDDL planners to obtain a solution to the WSC problem.

A three step process for the solution of WSC problems is presented in [7], which involves the translation of OWL-S descriptions and OWL ontologies to a PDDL domain and problem description, the solution of the planning problem through a planner, and the translation of the PDDL plan back to a composite OWL-S WS. Nevertheless, as the authors note, the focus of the paper is only on the first step, and the work presented is basically exploratory.

A novelty of our system is the fact that it will be open source and based on a publicly available online application. To our knowledge, there are currently no web-based systems supporting multiples phases of the WSC process available. YaWSA [8] was the only WSC system in the literature that allowed users to compose services from a web-based interface. However, it was simplistic and only implemented a WSC process, without offering a registry, or the ability to view and edit the WSs' descriptions online. Moreover, at the time of writing it was no longer available for public use.

A prototype web-based WSC system is described in [9], supporting WS browsing, the creation of composite services, service flow execution, and the generation of OWL-S descriptions used for describing their common process pattern instances. These instances are meant to bridge the gap between the users' requirements and the technical service descriptions, as the authors consider OWL-S to be insufficient and not abstract enough to achieve such a result on its own. However, a public link to a running demo of their implementation is not provided.

Finally, it should be noted that the recent bibliography [10] suggests a gap in the evaluation process of the current WSC systems. Not only is there no standard WS test set [4], but most approaches, especially the ones related to planning based techniques, simply evaluate their methodology on a single case study, without referring to quantitative criteria [11, 12, 13]. Only recently, however, a few approaches, such as [5, 6, 14], provided notable exceptions to this rule.

The most extensive evaluation results are provided in [4], which analyzes two artificial benchmarks with different encoding methods and planners, and measures the total runtime of the planner, as well as the number of search states and actions in the output plans. In [6] a single case study is presented, with a different number of WSs participating in the WSC experiments, and measuring the preprocessing, transformation (from OWL-S to PDDL) and planning time required. However, the use of only one planner is referenced, despite the possible use of two different ones, and the atomic WSs that comprise the composite one seem to be (mostly) hand-tailored by the authors, although entire domains of the OWL-S Service Retrieval Test Collection (OWL-S TC) [15] are used for the composition in general.

Finally, Kona et al. [14] present three detailed versions of a single use case scenario, each suited for a mode of their WSC algorithm, along with the IOPEs of the services that take part in the solution of the problem. The test collection used is also mentioned; a modified version of the 2006 WS-Challenge made to fit the authors' framework of choice, as well as various quantitative results regarding the experiments; the number of WSs participating in the WSC, the number of I/O parameters each WS had, and the preprocessing and query execution time needed to obtain a solution. In contrast to [4] and [6], however, the authors of [14] do not provide details regarding the machine that was used to run the experiments.

III. Proposed Approach

WSs' technologies are based on the idea of maximizing the reuse of loosely coupled components. As such, our view is that the systems implementing WS' functionalities should also be created with the same approach in mind and incorporate already freely available components as their sub-elements. Apart from the additional effort required to create a new component from scratch, such approaches have led to an abundance of applications and standards that only slightly differ from each other, while making the quantitative comparison of different systems difficult; this fact was illustrated in the previous section, and also demonstrated by various surveys relying only on qualitative criteria to review the available methodologies [10].

Our system supports various functionalities relating to different stages of WSC; the first one is the ability to store the service descriptions that will be used later in the discovery of suitable WSs in a registry. The core of the application is based on iServe [16], an open platform for publishing and discovering services. Specifically, we make use of its web-based application that allows users to browse, query and upload services, which, in our case, are semantically described in OWL-S. We have added an XML editor to the application, made several improvements to its interface and functionality, and populated the registry with version 4.0 of the OWL-S TC.

As aforementioned, the planning module will use NuPDDL for its purposes, as it is compatible with PDDL2.1, retaining most of it, including the handling of functions, conditional effects, and quantifiers; it is also capable of modeling non-deterministic action effects through the introduction of new keywords, such as *oneof* and *unknown*. Since the WSs in the registry are described semantically through OWL-S, a translation between the two languages must take place; we adopted an approach similar to [5, 6, 17] that imply that this conversion is straightforward, at least partially, and present their own mapping.

In [5], the OWL-S' ServiceProfile input parameters are converted to identically named ones of a PDDL action, and the *hasPrecondition* and *hasEffect* parameters to the precondition and effect of the action respectively; in [6], a similar approach is followed, with SWRL used to model the WSs' preconditions and positive effects, and RuleML to model their delete effects. However, [17] and [5] note the problematic conversion of non-physical knowledge from OWL-S inputs and outputs to PDDL. Both tackle the problem by introducing a new predicate in the PDDL domain; the first creates a predicate *agentKnows* with one argument that can either be bound to an input or an output parameter, while the second adds every output variable *X* to the world state through the introduction of an add-effect predicate *agentHasKnowledgeAbout(X)* (the same process is followed in an analogous manner for every input parameter).

After the translation, AI planning techniques can be used to generate the output plan/composite WS. We opt for the incorporation of a contingent planner, so as to generate plans that can cope with the most influential and likely contingencies, as composite WSs may fail to execute correctly for various reasons, such as the unavailability of an atomic WS involved in the plan, or simply because the output of their successful execution is not the expected one.

Our goal is not to develop a plan for every possible contingency, as the WSC domain has too many sources of uncertainty for such an approach to succeed. Instead, similarly to [18], we will produce a seed plan, examine it to determine significant or likely points of failure, and add a conditional branch to recover the plan's execution; this process will be repeated until we either reach a plateau or run out of time. As we cannot cope with every possible point of failure, a re-planning module will also be incorporated.

Finally, we will convert the NuPDDL plan back to an OWL-S (composite) WS, that is, create an OWL-S profile and its process description, without, however, providing a corresponding WSDL definition, in a fashion similar to that described in [14, 19]. In short, the profile description of the new composite WS will treat it as an atomic service with IOPEs, while the process model will be based on OWL-S control constructs that describe the way the WSs that compose the composite one interact with each other. The OWL-S API [20] that will be used to implement the conversion supports composite processes that use OWL-S control constructs, such as ⟨*Split-Join*⟩, and conditional constructs like ⟨*IfThenElse*⟩, which will be necessary to produce correct solutions to the use cases presented in the next section. Figure 1 illustrates our approach.

Our work is still in progress and the development of the web-based application is still in alpha version; as such it is not yet available publicly to users. However, a link to its current source code is available in [21].

IV. Evaluation of Results

As aforementioned, there is currently no standard WS test bed, concerning both the scenarios used to test the WSC process, and the WSs that take part in it. However, the recent trend of widespread use of OWL-S TC, as a test bed in the recent S3 contests [22], or in the recent literature [2, 5, 7], suggests its suitability for use in our evaluation experiments.

We believe that it is beneficiary to define specific use case scenarios in detail, as well as provide the actual WSs' descriptions that will be used. As such, we have designed three use case scenarios, each based on the WSs contained in a domain of OWL-S TC, and with an increasing amount of non-determinism and complexity than the previous one. In order to design useful test cases for our system, we made several minor modifications to the available WSs' descriptions and their relative ontologies, and also added a few descriptions to the collection, albeit similar to the ones already included in it. A full description of the use cases and the WSs they are based on can be found in [23].

The first use case is fully deterministic, allowing for the output of a fully serialized composite WS; it refers to a user who knows part of a movie title and wants to retrieve all the comedy films that exist with a similar title, along with their

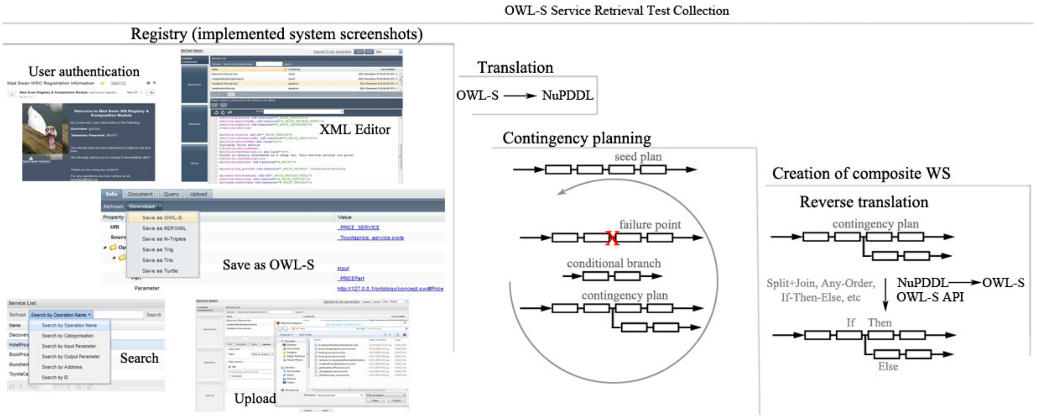


Figure 1. System overview (using a modified figure from [18]).

pricing information. The other two scenarios feature non-deterministic elements, such as preferences between types of products, or cases where a WS may have different outcomes. Particularly, the second one refers to an online bookstore user who wants to purchase a book with a preferred method of payment (a cheque, debit or credit card), with the WS having different outcomes depending on whether the book is in stock at the store or not. If it is available, the composite WS should add the book to the user's shopping cart, purchase it with the specified method of payment, and output information regarding it, such as its author. If, however, it is not in stock, no payment should be made, and no further information concerning it should be displayed to the client.

The third use case concerns the purchase of a camera; the user has a preference towards an analog SLR model, but is willing to settle for other ones if that one is not in stock. Apart from the addition of preferences, this scenario differs from the second one in that more than one sellers are assumed to exist, and the composite WS should check with all of them to determine if the item is in stock. As such, if a store is found that sells the analog SLR model and has it in stock, it should be added to the user's shopping cart. If it is not in stock, the search should continue for another store that sells it, and if one cannot be found, the process should be repeated, this time searching for the camera's compact version, or, if all else fails, for any camera available in stock.

Although the first two scenarios can be considered as special cases of the last one, it is important to showcase that the system can indeed cope with the generation of both sequential and conditional plans, with and without preferences. Moreover, the importance of the scenarios lies in that they exhibit that this particular test set can be used to produce meaningful use cases that can evaluate the capabilities of WSC approaches efficiently and in a manner that is reproducible and extensible.

V. Conclusion and Future Work

In this paper, we presented our ongoing work towards the implementation of an online WSC system that makes use of non-deterministic AI planning techniques and of already freely available WS-related components. Furthermore, we described in detail three use case scenarios that will be used to evaluate such systems, based on an existing WSs' test collection.

The fact that the final system will support various stages of WSC, as well as that it will be online and open source, is important, as at the moment, there is no other system with similar capabilities. Moreover, the scenarios enable us to test whether the proposed system can cope with the demands of WSC efficiently. In addition to this fact, it is our hope that they can be used by other WSC projects as a common test bed, since they provide detailed descriptions of the WSs involved, as well as their intended goals. Furthermore, they can be used by systems supporting either deterministic or non-detetermistic planning.

We expect that in the near future we will be able to demonstrate the first results of this effort through a publicly available online prototype.

Acknowledgment

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

References

[1] NuPDDL: Non-determinism and more in PDDL, http://mbp.fbk.eu/NuPDDL.html 03/08/2012

[2] E. Sirin, B. Parsia, D. Wu, J. Hendler, and D. Nau, "HTN planning for web service composition using SHOP2", *J. Web Semant.,* vol. 1, no. 4, Oct. 2004, pp. 377-396.

[3] E. Sirin, B. Parsia, and J. Hendler, "Template-based composition of semantic web services", Proc. *1st International AAAI Fall Symposium on Agents and the Semantic Web*, Nov. 2005.

[4] J. Hoffmann, P. Bertoli, M. Helmert, and M. Pistore, "Message-based web service composition, integrity constraints, and planning under uncertainty: a new connection", *J. Artif. Intell. Res*, vol. 35, May 2009, pp.49-117.

[5] M. Klusch, A. Gerber, and M. Schmidt, "Semantic web service composition planning with OWLS-Xplan", Proc. *1st International AAAI Fall Symposium on Agents and the Semantic Web*, Nov. 2005.

[6] O. Hatzi, D. Vrakas, M. Nikolaidou, et al., "An integrated approach to automated semantic web service composition through planning", IEEE Trans. Service Computing, April 2011, pp. 301-308.

[7] Y. Bo and Q. Zheng, "A method of semantic web service composition based PDDL", Proc. *IEEE International Conference on Service-Oriented Computing and Applications* (SOCA '09), Dec. 2009, pp. 1-4.

[8] A. Macdonald, "*Service composition with hyper-programming*". Technical Report, University of St Andrews, 2007.http://www.cs.st-andrews.ac.uk/~angus/docs/yawsa/final_report.pdf 03/09/2012

[9] X. Du, W. Song, and M. Munro, "Using common process patterns for semantic web services composition", Proc. *15th International Conference on Information Systems Development* (ISD'06), Sept. 2006.

[10] M. Chan, J. Bishop, and L. Baresi, "*Survey and comparison of planning techniques for web services composition*". Technical Report. University of Pretoria. 2007. http://polelo.cs.up.ac.za/papers/Chan-Bishop-Baresi.pdf 03/11/2012

[11] K. Chen, J. Xu, and S. Reiff-Marganiec, "Markov-HTN planning approach to enhance flexibility of automatic web service composition", Proc. *IEEE International Conference on Web Services* (ICWS'09), July 2009, pp. 9-16.

[12] D.V. McDermott, "Estimated-regression planning for interactions with web services", Proc. *Sixth International Conference on Artificial Intelligence Planning Systems* (AIPS '02), April 2002, pp. 204-211.

[13] S. McIlraith and T. Son, "Adapting Golog for composition of semantic web services", Proc. *8th International Conference on Principles of Knowledge Representation and Reasoning* (KR2002), April 2002, pp. 482-496.

[14] S. Kona, A. Bansal, M.B. Blake, and G. Gupta, "Generalized Semantics-Based Service Composition", Proc. *IEEE International Conference on Web Services (ICWS'08)*, Sept. 2008.

[15] SemWebCentral OWL-S Service Retrieval Test Collection, http://semwebcentral.org/frs/?group_id=89 02/15/2012

[16] C. Pedrinaci, D. Liu, M. Maleshkova, et al., "iServe: a linked services publishing platform", Proc. *Ontology Repositories and Editors for the Semantic Web Workshop at the 7th Extended Semantic Web Conference* (ORES '10), May 2010.

[17] P. Birchmeier, "*Semi-automated semantic web service composition planner supporting alternative plans synthesis and imprecise planning*". Diploma Thesis, University of Zurich, 2007. http://www.ifi.uzh.ch/pax/uploads/pdf/publication/1691/Birchmeier_Peter.pdf 03/12/2012

[18] R. Dearden, N. Meuleau, S. Ramakrishnan, D. Smith, and R. Washington, "Contingency planning for planetary rovers", Proc. *3rd International NASA Workshop on Planning & Scheduling for Space* (IWPSS '02), Oct. 2002.

[19] E. Ziaka, D. Vrakas, and N. Bassiliades, "Translating web services composition plans to OWL-S descriptions", Proc. *3rd International Conference on Agents and Artificial Intelligence* (ICAART '11), Jan. 2011, pp. 167-176.

[20] OWL-S API Introduction, http://on.cs.unibas.ch/owls-api/index.html 03/02/2012

[21] Application's source code, http://ai.uom.gr/gmarkou/Files/MadSwanSourceCode/ 03/11/2012

[22] S3 Contest: Retrieval Performance Evaluation of Matchmakers for Semantic Web Services, http://www-ags.dfki.uni-sb.de/~klusch/s3/html/2011.html 01/13/2012

[23] Markou, G.: "*Heracleitus II - WSC Use Case Scenarios*". Technical report. University of Macedonia, 2012. Available at http://ai.uom.gr/gmarkou/Files/Mad_Swan_Use_Case_Scenarios.pdf 04/26/2012

Towards Requirements Engineering for Mashups: State of the Art and Research Challenges

Vincent Tietz, Andreas Rümpel, Christian Liebing, and Klaus Meißner
Faculty of Computer Science
Technische Universität Dresden
Dresden, Germany
{*vincent.tietz,andreas.ruempel,christian.liebing,klaus.meissner*}@*tu-dresden.de*

***Abstract*—Mashups have become popular in the modern Web providing a lightweight development approach for mainly small and situational applications. Visual composition metaphors and loosely coupled widgets encourage the fast implementation of changing requirements. However, domain experts and mashup composers are poorly supported in expressing and formalizing their needs, leading to time-consuming and error-prone component discovery and composition. Methods and techniques of traditional requirements engineering (RE) are not transferable out of the box due to targeted user groups, isolated development phases, insufficient tool support, and different models. Therefore, we investigate characteristics of software engineering in mashup approaches compared to similar development paradigms revealing and discussing challenges when applying RE to mashups.**

***Keywords*-Web mashups; requirements engineering; software development process; application modeling; UI composition**

I. Introduction

Presentation-oriented *mashups* introduce the user interface (UI) as a new integration layer for component-based applications. They have become a prominent approach for the lightweight integration of distributed and decoupled Web resources. Originally, mashups have been developed by script-based assembly of heterogeneous application programming interfaces (APIs). However, incorporating UI fragments, the heterogeneity of mashup building parts and the composition effort increases. Thus, more powerful mashups are possible to be built at the expense of a simple development approach. Since there is no widely accepted understanding about a general mashup development process, also the manifestation of RE remains uncertain. Beside available work regarding traditional, component-based and Web-based RE, at least the necessity of a structured development process for enterprise mashups is recognized [1].

Mashups promise less expensive and faster application development of long tail applications integrating existing components and services customized by users. Apart from very simple, situational mashups, the explicit specification of functional and non-functional application requirements is essential. Driving factors are the reuse of pre-existing business processes or task models (cf. [2]), the growing application complexity, and business plans of service providers considering mainly non-functional requirements. However, the requirements elicitation and specification in mashups is neglected so far, making the quality-aware discovery and integration of components difficult. Moreover, missing model-based development approaches impede platform independence, adaptation, reusability, and maintainability.

To identify research challenges when applying RE to mashups in Section IV, we initially introduce the principles of RE, its core activities, and the characteristics of the application type *Web mashup* in Section II. Selected approaches are evaluated based on a catalog of criteria defined in Section III. Finally, Section V concludes this paper.

II. Principles

This section gives a brief overview of the principles of RE, the related core activities, and techniques. Further, we introduce Web mashups as our target application type.

A. Requirements Engineering

Software RE is the process of discovering the purpose of a software system by identifying and documenting stakeholders and their needs in order to achieve a satisfying software system for which it was intended [3]. Therefore, RE is multidisciplinary and human-centered, whereas the communication of requirements (implying readability, validity, and comprehensibility) and the management of requirements (implying traceability, searchability, and changeability) are critical success factors. The *development process* is considered as the instance of a process model defining *roles*, *artifacts*, and *activities*. The *technique* defines how to perform an activity, while the *method* combines both activities and techniques.

Traditionally, the first activity in RE is a *feasibility study* often coupled with a *risk analysis*. If the project effort is estimated to be adequate, the *elicitation and analysis* phase follows. There are a plenty of techniques available for requirements elicitation, such as *interviewing*, *brainstorming*, *analyzing existing documentation*, *data mining*, and *prototyping* [3]–[5]. In general, these techniques can be applied to different kinds of development processes. They are rather independent of the target application type applying social, psychological, and analytic strategies.

In contrast, the next activity, *elaboration and specification* of requirements, is intended to reveal the requirements in order to specify functional and non-functional product descriptions for subsequent development activities. The objective of structured requirements elicitation and specification is to achieve completeness and consistency, covering all requirements, and getting non-ambiguous specifications. The specified requirements constitute a contract, which forms the basis for later *validation* and iteratively making compromises within the *negotiation phase*. The *validation phase* is based on previously specified requirements. Depending on the domain and type of a software product, requirements change over time. Typically, *requirements change management* cares about additional or belated requirements or changes in previously taken decisions, also regarding potentially increasing costs or development effort.

The formalization degree of requirements artifacts affects the degree of automation for refinement or validation, e. g., using automated test cases, model checking or monitoring. To increase transferability and reusability, modeling techniques have been established to specify requirements in a formal way. One widely used modeling facility for object-oriented software is the Unified Modeling Language (UML). Formal specifications enable the refinement and transformation of models covering separate concerns of the development process and the final product [6]. The chance for model-driven refinement of computation-independent requirement models to technological independent conceptual models in the design phase is one major incentive for the application of model-driven RE. Ideally, the generation of executable code from these models leads to low-error software products fulfilling the specified requirements.

B. Web Mashups

Mashups evolved from simple data-driven aggregation of feeds to complex applications composing Web- and UI-based building parts. In general, tools like *Yahoo! Pipes*[1], *JackBe Presto*[2], and *IBM Mashup Center*[3] support the visual composition of technology-independent Web services, APIs, and UI components by dragging components on a canvas and wiring output and input channels in order to create new applications. Compared to other software systems, mashups are rather small applications with only few stakeholders, such as component developer, mashup composer, and mashup user. Regarding methods, patterns, and tools, simplicity and reusability are important demands, since mashup development by users with low programming skills, also referred to as *end user development*, is getting popular.

In general, mashup components can be considered as self-contained entities solving user tasks. Figure 1 shows an example Web mashup, which allows planning of a conference participation. In order to receive suggestions for routes of the public transportation services, a participant needs to define start and destination locations as well as corresponding temporal constraints. In addition, the user requires information about available hotels and the weather near the conference location. Therefore, the task *plan conference participation* is decomposed into *get travel info*, *get weather*, and *find hotel*. These tasks are supported by appropriate components (e. g., a map) that need to be selected and composed.

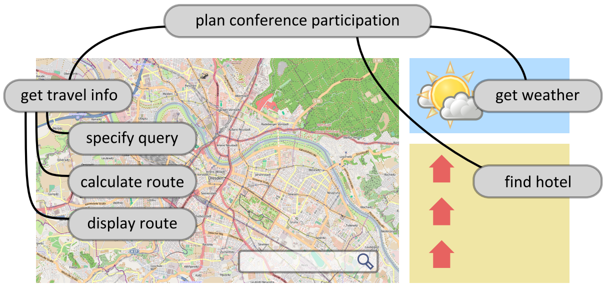


Figure 1. Conference participation mashup with tasks

Despite the simplicity of composition metaphors in current tools, the component discovery remains difficult. Searching is occasionally facilitated by keywords, interface descriptions, and community feedback. However, in the light of growing repositories and ambiguous tags, the identification of proper search criteria becomes an increasing challenge for inexperienced users. Further, mapping of component interfaces, e. g., originating from different providers, that should communicate within the same application, is not trivial. Thus, we argue that this should be addressed by a model-driven and semantics-based development approach.

III. State of the Art

To support RE, a wide range of development processes, methods, and tools have emerged. In the following, we especially evaluate model-driven development approaches for Web applications to review the state of the art and to reveal research challenges when applying RE to mashups.

A. Evaluation Criteria

As a foundation for our evaluation criteria, we adopted the evaluation framework in [7] for a survey of RE in Web- and service-based development approaches, facing mashup-relevant aspects regarding *requirements model*, component-aware *development process*, *model-driven development*, and *adequate tooling support*.

1) Requirements Model: Requirements, which should be provided by a software, are specified in a requirements model. Commonly, they are divided into non-functional, functional, and domain-related requirements [3]. *Functional requirements* define functions of a software system or parts of it. They are intended to accomplish calculations, data

[1] http://pipes.yahoo.com/
[2] http://jackbe.com/products/
[3] http://ibm.com/software/info/mashup-center/

manipulation, or other processing consisting of inputs, behavior, and outputs. From the user's point of view, mashups and their constituent components support the user in solving his or her specific tasks. Technical details and system characteristics are usually hidden by black-box components externalizing their functionality by their interfaces only. Therefore, we evaluate how functions (FR1), their sequence (FR2), their input and output (FR3), and the relationships of outputs and inputs (FR4) are specified.

Functional requirements are supported by *non-functional* requirements (NFRs), or *quality requirements*, which impose constraints on the design or implementation [3]. Product quality criteria in common development processes are mainly guided by the international standards for the evaluation of software quality ISO/IEC 9126 and ISO/IEC 25000. Internal and external metrics specify different quality criteria such as availability, efficiency, or security, defined in a quality model. If regarded, they are usually specified in poorly structured documents making it very hard to reuse them or automatically check and monitor their violations. Since NFRs imply a destination or *association point*, cf. [8], we evaluate, whether the method allows the specification of related objects in a suitable granularity. Such associated objects can be functional requirements, system artifacts such as components, resources, or the project context (NFR1). Further, specification possibilities of how a NFR is scaled or measured within the target system is evaluated as *integrability* or *operationalization* (NFR2).

Software is developed and used within an application domain. This includes, for example, the organization's structure, business rules, goals, tasks and responsibilities of its members, and the data that is needed, generated, and manipulated [3]. A comprehensive *domain model* provides an abstract description of the world in which an envisioned system will operate. It provides also knowledge structures and therefore allows reasoning on facts, as well as opportunities for reuse within a domain [3]. Because the domain model is an integral part of any requirements specification, we evaluate how domain models are supported describing data structures (DM1), roles and stakeholders (DM2) and how existing domain models can be reused (DM3).

2) Development Process: The development process needs to be adopted to the target user group in order to provide a lightweight and efficient development approach, wherein the knowledge about available components needs to be considered in the requirements phase. If requirements are identified at an early stage and components are chosen at a later stage, there is a bigger chance that components do not support the required features [9]. The main advantage of applying RE to mashups is that requirements can be matched to predefined components [10] and to support the development process by prototype generation, providing instant feedback. Therefore, we evaluate how the requirements phase is embedded in the whole development process, including the support of round-trip engineering (DP1) to automate model synchronization. Since we consider composite mashups, we evaluate whether the development process is component-based (DP2) and whether component recommendation is supported (DP3). Finally, we evaluate how the decomposition of requirements is supported (DP4). This also applies to NFRs, which are likely to be decomposed into NFRs or FRs.

3) Model-driven Development: We argue that model-driven development (MDD) [6] needs to be applied in order to benefit from traceability, usability, and platform independence in mashups [11]. Therefore, we evaluate existing work considering the provision of precise metamodels (MD1) and mappings and transformations into conceptual models based on the requirements model (MD2). Moreover, this implies some kind of component discovery (MD3) that may be facilitated by the use of semantic knowledge (MD4) [12].

4) Tool Support: Since mashup development is rather user- and prototype-driven, adequate tools are needed to support requirements elicitation and specification. Regarding the tool support we evaluate the provision of visual requirements modeling (TS1) and of managing development artifacts (TS2). We further try to identify the target user group of these tools to decide whether they are applicable to mashup composers (TS3) and how they are guided through the development process. To this end, we specified the following user groups: requirements engineer (RE), modeling expert (ME), software developer (SD), and end user (EU).

Since we regard model-driven development as a key demand and mashups as Web-based applications, we focus on corresponding development methods in our evaluation. Initially, we discuss traditional software engineering, as it provides the most comprehensive and mature approaches for RE and MDD. Finally, we consider current mashup approaches illustrating the lack of RE support.

B. Traditional methods and techniques

The notion *traditional software development* is used here to describe mature and well elaborated methods and techniques usually applied in industrial software projects and object-oriented development processes. Amongst industrial practitioners, the requirements elicitation phase is most often realized with the help of scenarios and use cases followed by focus groups and informal modeling [4]. Since object-oriented analysis is the most popular modeling method, we consider Rational Unified Process (RUP) [13] in conjunction with UML and related development tools, such as IBM Rational Software Architect (RSA)[4], as one representative of traditional software development methods.

1) Requirements Model: Regarding the requirements model, we observe that all functional (FR1–FR4) and non-functional requirements can be expressed with the help of use cases and textual supplements. Role assignments are

[4] http://www-01.ibm.com/software/awdtools/systemarchitect/

supported within the *NFR questionnaire* (NFR1) that can be enriched by textual impact descriptions (NFR2). Domain requirements (DM1) are described with the help of use cases and class diagrams. Stakeholder analysis is also supported by business analysis models (DM2). Principally, reuse of domain models is achieved by importing existing class diagrams. However, domain models tend to be application specific. Established repositories or standardized schemas as well as semantic mappings are usually not applied.

2) Development Process: RUP provides the disciplines business modeling, requirements, analysis and design, implementation, test, and deployment that all pass iteratively the phases inception, elaboration, construction, and transition. Within the business modeling discipline the system context is analyzed to capture structure and dynamics of the organization. Business requirements are captured by use case and analysis models that structure information about the organization and the relations to external stakeholders. In the requirements discipline, functional and non-functional requirements are refined by business use case models to system use cases and class diagrams. As round-trip engineering and the usage of component-based architectures are promoted practices in RUP we consider DP1 and DP2 as supported. However, the focus of round-trip engineering is on UML and Java. Components correspond rather to object-oriented artifacts that tend to be tightly coupled. The recommendation of components (DP3) based on the requirements analysis is not explicitly proposed. The decomposition of requirements (DP4) is partially supported by use cases, documents, and interactive refinement, because use cases are not intended for detailed functional decomposition.

3) Model-Driven Development: RUP provides no specific guidance on how to apply MDD, but offers an appropriate basis for it [14]. Although, the usually applied UML use cases provide a graphical notation, the metamodel is limited (MD1). Many aspects such as linking use cases by pre- and post-condition relations are not supported [15]. This leads to the use of text-based templates that do not provide formal precision, tend to be ambiguous and impede automated model transformations. Therefore, the transformation from computation independent model (CIM) to platform independent model (PIM) is limited (MD3). In general, the automatic discovery of components based on semantic knowledge (MD4) is not supported.

4) Tool Support: RSA is one example for the application of RUP for model-driven and object-oriented development. Since it provides visual modeling for use cases and domain models, we consider TS1 as supported. However, the tool is intended to be used by requirement engineers and modeling experts that should be experienced in using UML (TS3). IBM Rational Requirements Composer[5] is proposed to bring all stakeholders together for eliciting and managing requirements (TS2). Supported techniques comprise documents, storyboards, process diagrams, and use cases. Resulting use cases can be integrated in RSA providing traceability throughout the development process. However, detailed formal representations are rarely used and informal representations such as natural languages are still preferred.

C. Web Engineering Methods

In the past, the Web engineering community proposed several model-driven development methods. Prominent examples are OOHDM [16], OOWS [17], UWE [18], WSDM [19], and WebML [20]. The main purpose is the explicit support of Web-specific concerns such as navigation, presentation and personalization by providing conceptual models, and the generation of code. Although, the focus is mainly on the design phase, it is recognized that requirements analysis and specification need to be considered [5, 7].

1) Requirements Model: Regarding requirements analysis and specification usually existing techniques are adopted. For example, the functional requirements of OOHDM, UWE, and WebML are represented by use cases whereas functions (FR1), their sequence (FR2), input and output (FR3), and their relations (FR4) are mainly defined with the help of text-based templates. WSDM and OOWS propose textual templates and task descriptions such as ConcurTask-Tree (CTT) notation implying a lack of semantic clarity. Data requirements (DM1) are not explicitly supported by OOHDM, UWE, and WSDM [7]. Only OOWS uses information templates and WebML uses a data dictionary and entity relationship models to support data requirements explicitly. With the help of use cases or task models, all methods cover some kind of role specification (DM2). However, the reuse of existing models (DM3) is limited, because the development processes start with use cases leading to new and application-specific domain models.

2) Development Process: Regarding the development process, all methods apply some kind of requirements, design and generation phases, whereas round-trip engineering is not supported (DP1). The design phase is usually divided into conceptual, navigational, and functional modeling. Only OOHDM, its extension OOH4RIA [21], OOWS, and WebML are partly component-based (DP2) by using functionality that is provided by Web services. However, component recommendation is not available in any method (DP3). Web applications are usually generated with the help of templates or manually selected components. Further, the requirements decomposition (DP4) is completely supported in OOWS by using task trees and otherwise supported by use cases or textual descriptions.

3) Model-Driven Development: From the model-driven point of view (MD1–MD2), the specification of application requirements is not considered adequately in order to enable model-based transformations into conceptual models, because traditional Web engineering methods, except OOWS

[5] http://www-01.ibm.com/software/awdtools/rrc/

and UWE, do not provide a metamodel for requirements analysis [7]. Therefore, the automatic transformation to subsequent artifacts is not available. Since there is no consistent use of components, their discovery (MD3) using semantic knowledge (MD4) is not supported sufficiently.

4) Tool Support: OOHDM has been supported by HyperDE[6], but the development seems to be discontinued since 2009. In principal, UWE can be used in any tool supporting UML stereotypes, but there is also the specialized tool MagicUWE[7] available. WebML is supported by WebRatio[8] that allows business process modeling, application modeling, generation, and deployment. OOWS provides a CASE tool based on the Eclipse platform [7]. There is no dedicated tool for WSDM available, however CTTE[9] can be applied for task modeling. Overall, only WebRatio and MagicUWE appear to be maintained continuously. Tool support for managing created models or components (TS2) is currently not available. Regarding the target user group (TS3), users need to have knowledge about the associated requirement and modeling techniques. Therefore, model-driven Web engineering tools require similar skills as in traditional methods.

D. Service-Based Engineering

Building applications upon Web services provides simplification of application development by reducing the need for specific code. However, the black-box character of services impedes the prediction of the whole application behavior, which makes quality handling difficult [9] and leads to dependencies on vendors and less flexibilities in requirements. Various approaches for service composition at technical data integration level, e. g., BPEL, have been proposed, whereby the composition is realized using structured programming constructs without considering the interaction with users. To integrate human tasks, BPEL4People has been introduced. However, the problem of service composition at presentation layer is still not supported adequately. The most prominent approaches in this context are ServFace [22], MARIA [23], Achilleos et al. [24], and Tsai et al. [25].

1) Requirements Model: In general, these approaches do not support specific requirements modeling (FR1–NFR2), since their development is achieved by modeling directly on functional interfaces. They are represented by visual service front ends, generated dynamically. Only MARIA supports some kind of RE by task modeling in order to specify functions (FR1), their sequence (FR2), and input and output (FR3) including their relations (FR4). Analogous to Web engineering approaches, service-based engineering does not support entities and relations (DM1) explicitly. In MARIA, roles and stakeholders (DM2) can be associated with different task trees. Finally, the reuse of existing models (DM3) is partly supported in MARIA, while in other approaches the development process starts from the scratch at any time.

2) Development Process: The development process provides mainly design and generation phases. Round-trip engineering (DP1) is not supported, since the approaches are based on only one application model or apply sequential processes, whereby the models involved cannot be synchronized. All approaches are partly component-based due to the use of Web services. Because in general the UIs are generated dynamically, only Tsai et al. support UI services (DP2) and provide recommendation (DP3). Finally, except MARIA that uses task models, requirements decomposition (DP4) is not supported since RE is missing at all.

3) Model-Driven Development: Due to the overall lack of requirements models (MD1), all approaches do not support the transformation of user requirements into conceptual models (MD2). Solely MARIA uses CTTs to specify requirements in a first step, which are transformed into conceptual models. Regarding the component discovery (MD3), all approaches provide Web service repositories to allow a simple search by keywords. Since the approach of Tsai et al. is based on UI services, the corresponding registry supports discovery using semantic knowledge (MD4).

4) Tool Support: Regarding the tool support (TS1) and target user group (TS2), ServFace offers a visual and Web-based authoring tool[10] mainly for end users. MARIAE[11] supports task modeling in the context of MARIA and is mainly intended for requirements engineers. The other two approaches also address application developers with appropriate skills, whereby only the tool of Achilleos et al. provides visual modeling. The management of artifacts (TS2) is limited to Web service components and solely in Tsai et al. UI services may be managed within a repository.

E. Mashup Development Methods

Originally, mashups have been developed by manual, script-based integration of heterogeneous APIs. Addressing non-programmers, mashup tools like *Yahoo! Pipes* and *IBM Mashup Center* have emerged to support the visual composition of technology-independent Web services, APIs, and UI components. Since these tools do not provide model-based composition, we consider model-based integration platforms, providing component and composition models such as mashArt [26] and CRUISe [27].

1) Requirements Model: In general, there is no explicit requirements model in mashups available. Usually, the mashup composer is able to search for components by keywords or other criteria without being supported explicitly in expressing requirements. At the level of implementation, composition models or integration templates can be

[6] http://www.tecweb.inf.puc-rio.br/hyperde/wiki

[7] http://uwe.pst.ifi.lmu.de/toolMagicUWE.html

[8] http://www.webratio.com

[9] http://giove.isti.cnr.it/ctte.html

[10] http://www.servface.org/index.php?view=article&id=117

[11] http://giove.isti.cnr.it/tools/MARIAE/home

considered as requirements for automated matching purposes. Thus, functional specification is only achieved by manually selecting suitable mashup components using their interface descriptions (FR1, FR3). Sequences and workflow-structuring elements can be modeled using application-internal communication paradigms (FR2, FR4). However, this blends into design and implementation activities and does not produce user requirements directly (FR1–FR4). Although first mashup-specific quality modeling approaches exist [28] and prototypically extend mashArt, non-functional *user requirements* specification is not supported (NFR1, NFR2). Data structures (DM1) are predefined and limited to mashup component interface type definitions. Typically, a mashup application is made for one specific consumer role or end user, thus yielding to few variability in roles (DM2). The tripartite mashup role model applies mashup component developer, mashup composer, and end user. In existing mashup approaches, prevalent models, such as business processes, task models, or domain models cannot be reused as input for mashup development (DM3), but would be possible by adding model transformation engines to provide richer input facilities.

2) Development Process: By adding and rewiring new mashup components, a composition can be easily extended or changed, allowing evolutionary development. However, round-trip engineering is not supported due to single application models. The existing approaches are fully component-based (DP2), whereby in CRUISe and mashArt a recommendation mechanism (DP3) supports the integration of suitable mashup components. Using flat hierarchies, requirements can only be matched to component capabilities or the application logic provided by composing them. Thus, requirements decomposition is very limited (DP4).

3) Model-Driven Development: In CRUISe, formalized domain models are used by semantically enriched mashup component descriptions [12] (MD1). However, the model-based requirement specification and derivation of conceptual models is neglected so far. Different model transformations (MD2) are available in CRUISe, but there is no transformation or mapping of user requirements to conceptual models. Similar approaches are not able to cover them as well. Discovery (MD4) is supported via a component repository making use of semantic component descriptions [12]. It is used for interface query for component implementations facilitating integration and exchange (MD3).

4) Tool support: Visual composition tooling is currently not supported in CRUISe, but mashArt provides a user-scoped *mashArt editor* [26] (TS1). Beside the management of mashup components in a repository, included Web-based resources are managed via URI addressing and may be heterogeneously hosted. There is no unified facility of hosting composite applications (TS2). At least people with medium modeling skills and knowledge of component communication paradigms are required throughout all model-based mashup development approaches without appropriate visual tooling support (TS3). To this end, model-based mashup development platforms require advanced modeling skills, while providing fast application results by selecting pre-existing mashup components.

IV. Discussion and Research Challenges

The results of our evaluation are summarized in Table I, whereas it is obvious that all mashup approaches do not consider any kind of RE. In fact, the mashup composer is still constrained to decompose requirements mentally or with the help of other methods. This impedes component recommendation and composition based on user requirements. Apart from that, model-driven Web engineering methods propose several techniques, such as use cases, text templates, and task trees to document requirements. However, over 30 % of practitioners state that they do not model requirements at all [4]. The avoidance of formal representations during the requirements specification phase emphasizes their opinion, that the formalization effort does not pan out. Therefore, we argue that specialized RE for mashups is needed. Based on this, we identified the following main research challenges to apply RE to mashups:

1) How can essential mashup-specific requirements be sufficiently described by formal models? The requirements for mashups differ from the requirements in other development processes, because mashups are built upon components at a characteristic level of granularity and incorporate user interface parameters. In contrast to existing methods and techniques, semantic clarity needs to be achieved in order to enable model-driven and recommendation-based development support. This implies the shift of pragmatic mashup composition to semantic mashup composition and the incorporation of ontologies. While quality requirements are covered in traditional RE methods mainly in textual form, they are hardly observed in mashup approaches. At least, many non-functional requirements with great importance in distributed mashup application scenarios, such as performance constraints or communication security restrictions, can be formalized and measured very well. Therefore, NFRs should be part of the requirements model, specified together with violation consequences, making use of formalized model connections. Also, workflow aspects need to be integrated in such a requirements model to support more complex scenarios in enterprise mashups.

2) How can a requirements model support the composition phase? In fact, mashup development is currently reduced to the design and implementation phase. However, having appropriate requirements models available, the composition can be significantly improved by applying MDD and requirements-based component recommendation. The benefit of MDD for traditional software systems is relatively low, because the refinement steps are still performed manually. In contrast, mashups allow for aligning specified requirements

Table I
REQUIREMENTS ENGINEERING FOR MASHUPS AND RELATED DEVELOPMENT PARADIGMS: EVALUATION RESULTS

		RUP	OOHDM	OOWS	UWE	WSDM	WebML	OOH4RIA	Servface	MARIA	Achilleos	Tsai	CRUISe	mashArt
Requirements Model														
FR1	Functions	Use cases, Scenarios	Use cases, Text templates	Task charact. templates	Functional use cases	Natural language, Task models	Use cases, Text templates	No	No	Task models	No	No	No	No
FR2	Sequence of functions	Scenarios, Activity diagrams	Use cases, Text templates	Activity diagrams	Activity diagrams	Natural language, Task models	Activity diagrams	No	No	Task models	No	No	No	No
FR3	Input and output	Scenarios, Use cases (Pre- and post-conditions)	Use cases (Pre- and post-conditions), Text templates	Task characterization templates	Use cases (Pre- and post-conditions)	Natural language, Task models	Use cases (Pre- and post-conditions)	No	No	Task models	No	No	No	No
FR4	Relation between output and input	Scenarios, Activity diagrams	Use cases (Pre- and post-conditions), Text templates	Task characterization templates	Use cases (Pre- and post-conditions)	Natural language, Task models	Use cases (Pre- and post-conditions)	No	No	Task models	No	No	No	No
NFR1	Assoc. point model.	Questionnaire	No	No	No	No	No	No	No	No	No	No	No	No
NFR2	Operationalization	Partly (Textual impact descr.)	No	No	No	No	No	No	No	No	No	No	No	No
DM1	Entities and relations	UML class diagram, Business object model, Glossary	Information types	Information Templates	Not explicitly (Content elements)	Not explicitly (Textual descriptions)	Data Dictionairy, Text Templates	Domain models	Not explicitly	Partially (No domain modelling)	Not explicitly	Not explicitly	Predefined by component interface	Predefined by component interface
DM2	Roles and stakeholders	Use cases, Scenarios, Documents	Access capabilities	Task Descr., Interaction Points	Not explicitly (Roles in use cases)	Audience Modeling and Classification	User groups	No	No	Collaboration Tree	No	No	No	No
DM3	Reuse of existing models	Yes	Partly (Design Patterns)	No	No	No	Partly (BP models)	Partly (Domain model)	No	Partly (Task model)	Partly (PML model)	No	No	No
Development Process														
DP1	Round-trip engineering	Partly (UML, Java)	No	No	No	No	No	No	One model	No	No	No	One model	One model
DP2	Component-based development	Partly (Object-oriented)	Partly (Object-oriented)	Partly (Service widgets)	No	Partly (Service model)	Partly (By extension)	Yes	Partly (Web services)	Partly (Web services)	Partly (Web services)	Yes	Yes	Yes
DP3	Component recommendation	No	No	No	No	No	No	No	Partly (UI widgets)	No	No	Yes (UI services)	Yes	Yes
DP4	Requirements decomposition	Partly (Use cases)	Partly (Use cases)	Task trees	Partly (Use cases)	Partly (Textual descriptions)	Partly (Use cases)	No	No	Task trees	No	No	No	No
Model-Driven Development														
MD1	Precise req. metamodel	Partly (Use cases)	Partly (Use cases)	Yes	Yes (WebRE)	No	Partly	No	No	No	No	No	No	No
MD2	Mappings and transformations	Partly (Guidelines)	No	Yes	Yes	No	Partly (Guidelines)	Yes	Yes	Yes	Yes	Yes	No	No
MD3	Component discovery	No	No	No	No	No	No	No	Web service repository	Web service repository	No	UI service repository	Component repository	Component repository
MD4	Usage of semantics	No	No	No	No	No	No	No	No	No	No	UI ontology	SMCDL	No
Tool Support														
TS1	Visual modeling	Yes	No	Yes	Yes	Partly	Yes	Yes	Yes	Yes	Yes	No	No	Yes
TS2	Artifact management	Yes	No	No	No	No	No	No	No	No	No	Yes	Components	Components
TS3	Target user group	RE, ME	RE, ME	RE	RE, ME	RE	RE, ME	ME	EU	RE, ME	ME	SD	ME	EU

with those pre-specified in components. However, since mashups are usually fast developed for situational needs, further development steps need to be robust concerning inconsistent and incomplete requirements. Additionally, round-trip engineering is a necessary prerequisite to make model-driven application development practicable that is currently not provided in any Web-based engineering approach.

3) How can the mashup composer be supported in identifying and formalizing his or her needs? In general, mashups are intended to be created by end users or domain experts. Therefore, appropriate tools are needed to hide complexity and to guide mashup composers in expressing their requirements. This implies visual modeling of requirements and compositions as well as fast application generation. In fact, tools should provide concrete requirements activities which blend into the existing design activities, e. g., by drawing required tasks and dynamic refinement. At the same time, semantic modeling needs to be supported easily, incorporating established ontologies. Application generation and fast prototyping needs to be provided to support users in reviewing and adjusting their requirements.

V. CONCLUSION

This paper presents an evaluation framework consisting of several criteria to analyze RE in software development paradigms, facing relevant aspects of building *Web mashups*. To this end, we analyzed prominent approaches in detail that may be used to develop our target application type. By applying our criteria, it turned out that each of the development approaches provides a distinctive development method, while none is supported at all in existing mashup platforms. However, RE is crucial for mashup development as well to achieve an efficient and requirements-aware development. To substantiate this, we identified three main research challenges. We propose the specification of requirements in formal models to increase reusability, the establishment of a requirements model to support the composition phase and finally an adequate visual authoring tool that supports the mashup composer in identifying and formalizing his or her needs. Thus, we are convinced that addressing these challenges will yield to a more efficient, user-friendly, and quality-aware mashup development process.

Acknowledgment

The work of Vincent Tietz is funded by the European Social Fund (ESF), Free State of Saxony and Saxonia Systems AG (Germany, Dresden), filed under ESF-080939514. Christian Liebing is funded by the ESF and Free State of Saxony (Germany), filed under ESF-100088145.

References

[1] W. Ketter, M. Banjanin, R. Guikers, and A. Kayser, "Introducing an agile method for enterprise mash-up component development," in *IEEE Conf. on Commerce and Enterprise Computing*, Jul. 2009, pp. 293–300.

[2] V. Tietz, S. Pietschmann, G. Blichmann, K. Meißner, A. Casall, and B. Grams, "Towards task-based development of enterprise mashups," in *Proc. of the 13th Intl. Conf. on Information Integration and Web-based Applications & Services*, Dec. 2011.

[3] B. Nuseibeh and S. Easterbrook, "Requirements engineering: a roadmap," in *Proc. of the Conf. on The Future of Software Engineering*, ser. ICSE '00, 2000, pp. 35–46.

[4] C. Neill and P. Laplante, "Requirements engineering: the state of the practice," *Software, IEEE*, vol. 20, no. 6, pp. 40–45, Nov.–Dec. 2003.

[5] J. Escalona and N. Koch, "Requirements engineering for web applications – a comparative study," *Journal of Web Engineering*, vol. 2, no. 3, pp. 193–212, 2004.

[6] B. Selic, "The pragmatics of model-driven development," *IEEE Softw.*, vol. 20, no. 5, pp. 19–25, 2003.

[7] P. Valderas and V. Pelechano, "A survey of requirements specification in model-driven development of web applications," *ACM Trans. on the Web*, vol. 5, no. 2, pp. 1–51, May 2011.

[8] M. Kassab, O. Ormandjieva, and M. Daneva, *A Metamodel for Tracing Non-functional Requirements*. IEEE Computer Society, Apr. 2009, vol. 7, pp. 687–694.

[9] G. Kotonya, J. Hutchinson, and B. Bloin, *Service-oriented software system engineering: challenges and practices*. Idea Group, 2005, ch. A Method for Formulating and Architecting Component and Service-oriented Systems, pp. 155–181.

[10] V. Tietz, G. Blichmann, S. Pietschmann, and K. Meißner, "Task-based recommendation of mashup components," in *Proc. of the 3rd International Workshop on Lightweight Integration on the Web*. Springer, Jun. 2011.

[11] S. Pietschmann, "A model-driven development process and runtime platform for adaptive composite web applications," *International Journal On Advances in Internet Technology*, vol. 4, 2010.

[12] S. Pietschmann, C. Radeck, and K. Meißner, "Semantics-based discovery, selection and mediation for presentation-oriented mashups," in *Proc. of the 5th International Workshop on Web APIs and Service Mashups*, 2011.

[13] P. Kruchten, *The rational unified process: An introduction*. Pearson Education Limited, 2004.

[14] A. Brown and J. Conallen, "An introduction to model-driven architecture (Part III): How MDA affects the iterative development process," http://www.ibm.com/developerworks/rational/library/may05/brown/, 2005.

[15] G. Génova, J. Llorens, P. Metz, R. Prieto-Díaz, and H. Astudillo, "Open issues in industrial use case modeling," in *UML Modeling Languages and Applications*, ser. Lecture Notes in Computer Science. Springer, 2005, vol. 3297, pp. 52–61.

[16] D. Schwabe, G. Rossi, and S. D. J. Barbosa, "Systematic hypermedia application design with OOHDM," in *Proc. of the 7th ACM Conf. on Hypertext*, ser. HYPERTEXT '96. New York, NY, USA: ACM, 1996, pp. 116–128.

[17] J. Fons, V. Pelechano, M. Albert, and Óscar Pastor, "Development of web applications from web enhanced conceptual schemas," in *Conceptual Modeling – ER 2003*, ser. Lecture Notes in Computer Science, vol. 2813, 2003, pp. 232–245.

[18] A. Kraus, A. Knapp, and N. Koch, "Model-driven generation of web applications in UWE," in *Proc. of 3rd Intl. Workshop on Model-Driven Web Engineering*, 2007.

[19] O. Troyer and C. J. Leune, "WSDM: a user centered design method for web sites," in *Proc. of the 7th Intl. WWW Conf.*, 1998, pp. 85–94.

[20] S. Ceri, P. Fraternali, and A. Bongio, "Web modeling language (WebML): a modeling language for designing web sites," *Comput. Netw.*, vol. 33, pp. 137–157, June 2000.

[21] S. Melia, J. Gomez, S. Perez, and O. Diaz, "A model-driven development for GWT-based rich internet applications with ooh4ria," in *Intl. Conf. on Web Eng.*, 2008, pp. 13–23.

[22] M. Feldmann, T. Nestler, K. Muthmann, U. Jugel, G. Hübsch, and A. Schill, "Overview of an end-user enabled model-driven development approach for interactive applications based on annotated services," *Proc. of the 4th Workshop on Emerging Web Services Technology*, pp. 19–28, 2009.

[23] F. Paternò, C. Santoro, and L. D. Spano, "Exploiting web service annotations in model-based user interface development," in *Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems*, ser. EICS '10. New York, NY, USA: ACM, 2010, pp. 219–224.

[24] A. Achilleos, G. M. Kapitsaki, and G. A. Papadopoulos, "A Model-Driven Framework for Developing Web Service Oriented Applications," in *ICWE'11*, 2011.

[25] W.-T. Tsai, Q. Huang, J. Elston, and Y. Chen, "Service-Oriented User Interface Modeling and Composition," *IEEE Intl. Conf. on e-Business Engineering*, pp. 21–28, 2008.

[26] F. Daniel, F. Casati, B. Benatallah, and M.-C. Shan, "Hosted universal composition: Models, languages and infrastructure in mashart," in *Conceptual Modeling – ER 2009*. Springer, 2009, vol. 5829, pp. 428–443.

[27] S. Pietschmann, V. Tietz, J. Reimann, C. Liebing, M. Pohle, and K. Meißner, "A metamodel for context-aware component-based mashup applications," in *Proc. of the 12th Intl. Conf. on Information Integration and Web-based Applications*, 2010.

[28] M. Picozzi, M. Rodolfi, C. Cappiello, and M. Matera, "Quality-based Recommendations for Mashup Composition," in *Proc. of the ComposableWeb Workshop*, 2010.

Online Internet Communication using an XML Compressor

Tomasz Müldner
Jodrey School of Computer Science Acadia University
Wolfville, B4P 2A9 NS, Canada
e-mail: Tomasz.muldner@acadiau.ca

Jan Krzysztof Miziołek
IBI AL University of Warsaw, Poland
e-mail: jkm@ibi.uw.edu.pl

Christopher Fry
Jodrey School of Computer Science Acadia University
Wolfville, B4P 2A9 NS, Canada
e-mail: chrisfry99@gmail.com

Abstract—Online communication and various other Web applications, such as collaborative systems using XML as a data representation often suffer from performance problems caused by the verbose nature of XML. In this paper, we present an XML-conscious compressor designed to alleviate these problems, by using it online and evaluating queries using lazy decompression. Our XML compressor not only decompresses the data whenever enough data are available, but it also compresses them online, it is updateable (i.e., it works with dynamic XML documents), and its implementation can be parallelized thereby significantly increasing performance on multi-core machines.

Keywords - Internet communication; XML; compression.

I. INTRODUCTION

Online communication is increasingly using the eXtensible Markup Language (XML) [1] as a data format. Unfortunately, the XML markup results in increased size of this representation, often by as much as ten times as large as alternative formats. This overhead is particularly concerning for communications involving large data sets and for memory-constrained devices participating in online communication. For applications passing communicating over the Internet, the network bandwidth is the main bottleneck and this is why decreasing the size of the information passed is essential. There has been considerable research on XML-conscious compressors, which unlike general data compressors can take advantage of the XML structure, e.g., [2], [3], [4]. Most recently, there has been research on queryable XML compressors for which queries can be answered using lazy decompression, i.e., decompressing as little as possible, see, e.g., [5], [6]. Also, there has been research for updateable XML compressors, for which updates can be saved without full decompression, see, e.g., [7], [8]. Online XML compressors are typically defined as compressors, which decompress chunks of compressed data whenever possible rather than doing it offline when the entire compressed file is available, see [9], [10]. These compressors are particularly useful for Internet applications, used on networks with limited bandwidth. Clearly, for a compressor to be online means that it must have only one pass through the document is required to compress it. In this paper, we present an online compression based on XSAQCT, an XML compressor, see [11]. There are other online compressors, e.g., TREECHOP [12], but XSAQCT has a number of distinctive features, as it is queryable using lazy decompression (i.e., with minimal decompression) and updateable [7], and finally it can be parallelized to execute faster on multi-core machines [13]. Various possible educational applications of XSAQCT are described in [14]. Similarly to TREECHOP, XSAQCT supports both the compression where the decompressor's output is exactly the same as the original input (including the white space), and generating a canonicalized [15] XML document.

Contributions. We present a novel online XML compressor suitable for improving online communication on Internet. Our initial experiments indicate that XSAQCT's performance is comparable to TREECHOP, but unlike TREECHOP, XSAQCT not only decompresses the data whenever enough data are available, but it also compresses them online, which is essential for the case of a network node N1 receiving streamed XML data from one or more sources, which are to be stored in a compressed form. In such cases, instead of waiting for the entire set of XML data, N1 may compress incoming data online thereby increasing the efficiency of the compression. In addition, XSAQCT is updateable (i.e., it works with dynamic XML documents) and its implementation can be parallelized thereby significantly machines. increasing performance on multi-core This paper is organized as follows. Section II gives a short introduction to the design and functionality of the previous version of XSAQCT, which is offline, and Section III describes its current extension, i.e., the online XSAQCT. Section IV describes applications of XSAQCT to an online communication, and finally, Section V provides conclusions and describes future work.

II. OUTLINE OF OFFLINE XSAQCT

Given an XML document D, we perform a single SAX (specifically using Xerces [16]) traversal of D to encode it, thereby creating an annotated tree $T_{A,D}$ n which all similar paths (i.e., paths that are identical, possibly with the exception of the last component, which is the data value) are merged into a single path and each node is annotated with a sequence of integers; see Fig. 1. When the annotated tree is being created, data values are output to the appropriate data containers. Next, $T_{A,D}$ is compressed by writing its annotations to one container and finally all containers are compressed using a selected back-end compressors, e.g., gzip [17]. For more details on XSAQCT, see [11] and [7].

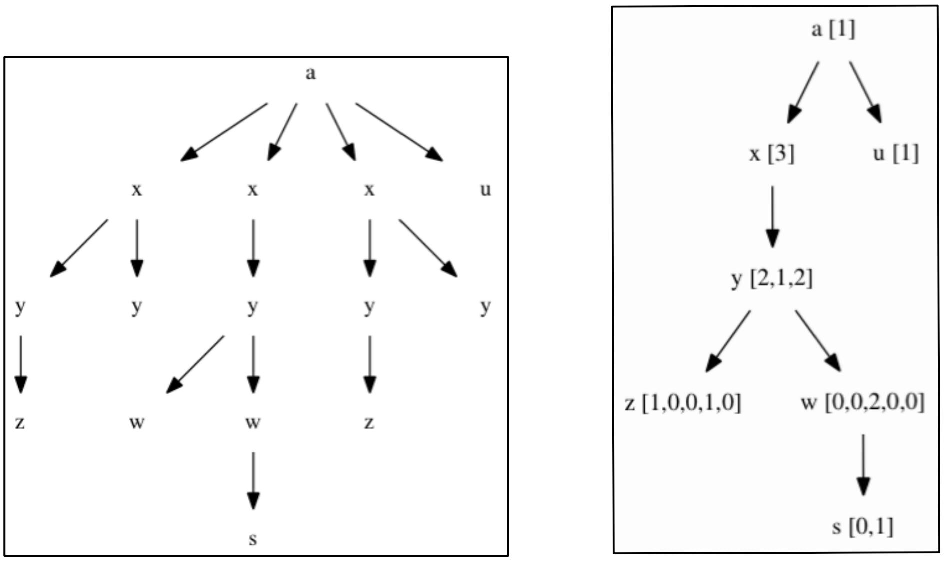


Fig.1 XML document (a) and its annotated tree (b)

III. ONLINE XSAQCT

A. Notations and Assumptions

In this version, we assume that a leaf of an XML tree stores exactly one text child. By SN we denote a sending node and by RN we denote a receiving node. SN and RN communicate using message passing; here SN is a producer and RN is a consumer using receive(packet); a synchronization is taken care of by these procedures. A packet may have binary contents or it can be a sequence of integer values. By the skeleton tree T_D we denote the tree labeled by tag names, and by we denote an annotated tree. By ANN we denote the sequence of all annotations. Annotations for a node of $T_{A,D}$ may be stored with this node, or the node may store a (logical) pointer to ANN (e.g., the offset within ANN). We assume that an annotated tree $T_{A,D}$ is implemented so that following functions are available:

- Node ADD_RC(Node n, Tag p, annotation a) creates and returns a new rightmost child of n with the tag p and the annotation a;
- Node create_Root(Tag p) creates a new root with tag p;
- Node get_LC(Node n) returns the leftmost child of n or a null node;
- Node get_RS(Node n) returns the right sibling of n or a null node;
- bool function is_Text(Node n) returns true iff n is a special tree node to store text;
- Node get_Parent(n) returns the parent of n; text get_Tag(n) returns the tag of n;
- text get_Text(n) returns the only text child of n.

We also assume that a data structure Path stores tags or text value, with the operations append_Path(Path p, Node n) which appends n to the path p, clear_Path(Path p) which sets the path p to empty, length_Path(Path p) which returns the length of p, and set_Path(Path p) which stores length_Path(p) as the first element of p. Finally, we use the following notations:

- a(n) annotation of the node n
- a(n)+=j increase the last annotation of n by j
- a(n)+=“,0” add “, 0” to the annotation of n; e.g. if a(n)=[1] then it becomes [1,0]
- $[0^{a(m)},1]$ if a(m) is [1], then $[0^{a(m)},1]$ is [0,1] otherwise $[0^{a(m)},1]$ is [0,…,0, 1] where $0^{a(m)}$ is the sum of all annotations in a(m), minus 1; e.g., if a(m) = [2,1], then $[0^{a(m)},1]$ is [0, 0, 1].

B. Online Compression

SN parses XML data using the SAX parser, and sends packets to RN, which first creates an annotated tree (as described below) and then follows the compression process as in XSAQCT [11]. At the same time, the parser creates a dictionary of tags. Each packet is of the form: (integer k, followed by N indices into the dictionary, followed by the uncompressed text) where

N>=0. We assume that when the SAX parser terminates (i.e., completes parsing) it sends the packet (-2, dictionary). Therefore, RN creates an annotated tree labeled by indices rather than tags. For the sake of readability, in the description provided in this paper, we consider sending and receiving tags rather than indices but our implementation operates on indices.

```
int k = -1;
Path p;
//initially stores only the tag of the root of the XML tree
void SN_send(Node n) {
        if (n is a leaf) {
              // a leaf must have a text child
                append_Path(p, getText(n));
                set_Path(p);
                send(p);
                clear_Path(p);
                k=0;
        } else
                for every child m of n {
                        append_Path(p, getTag(m));
                        SN_send(m);
                        k++;
                }
} // SN_send()
```

Now, we'll explain the actions executed by RN. By the *leftmost path* of the labeled tree T rooted at node n_1 we mean a path of labels $(p_1,\ldots,p_k)$ s.t. get_Tag(n_1) = p_1, and for i = 1,…,k-1 get_LC(n_i) = n_{i+1}, get_Tag(n_i) = p_i, get_LC(n_k) is null. The *first time* send() is called, it will send a packet (-1, the path of the leftmost path rooted at the root of the tree). At this time the value of the *current node* c will be set to n_k. To explain the meaning of sending the consecutive packet $(k, p_1,\ldots,p_N)$ let's assume that SN_send() is visiting nodes (which are initially non-visited) in a dfs-order and c is current node. Then the value of k is found as follows. Let n_1, …,n_k be the shortest path of nodes s.t. n_1 = c, for i=1,…k-1 get_Parent(n_i) = n_{i+1} and there exists a child m of n_k which has not been visited. Next, let $m_1,\ldots,m_N$ be a path of nodes s.t. m_1=m, for i=1,…,N-1 m_{i+1} is the leftmost unvisited child of m_i, m_N has only a text child. Then, for i=1,…,N get_tag(m_i)=p_i.

```
void RN_receive(Node n) {
  bool flag;
  Node c; // current node
  Node m;
  Text t;
  receive(k, p1,…,pN , t);
  if(k==-1) { // initialization
      c = create_Root(p1);
      for(i=2; i<N; ++i)
              c = ADD_RC(c, pi, [1]);
  }
  while (true) { // until the final packet
       receive(k, p1,…,pN , t);
       if(k == -2)
           return; // done
          // move current based on the value of c
         for(i=1; i<=k; ++i) // set the current
           c = get_Parent(c);
        // check every tag in the received path
        for(i=1; i<=N; ++i) {
            flag = false;
            for (every child m of c)
                if(get_Tag(m) == pi) {
                    a(m)+=1;
                    c = m;
                    flag = true;
                    for (every child m of c)
                        a(m) += ",0";
                    break;
                } // end of if
            // for every child
            if(!flag)
                c = ADD_RC(c, pi, $[0^{a(c)},1]$);
            add text to the container for c;
        } // for i=1…
    } // while(true)
} // RN_receive()
```

C. Online Decompression

We assume that the sending node SN can decompress all annotations, restore the skeleton tree and send it to RN, then re-annotate it as well as run a procedure SN_dfs(AnnotationTreeNode) shown below. As far as the receiving node RN is concerned, we assume that it can run a procedure RN_restore(SkeletonTreeNode) shown below. We also assume that RN implements the AA Abstract Data Type (ADT), which stores sequences of annotations with the following operations (initially, the annotations for every node are un-initialized):

- void AA_delete(Node n) removes the first element of the annotations for n
- void AA_store(Node n, sequence of integers seq) stores seq as the annotations for n
- void AA_init(Node n) initializes the annotations for n
- bool AA_isInit(Node n) returns true iff the annotation for n has been initialized
- int AA_getFirst(Node n) returns the first element from the annotations for n
- AA_get_Text(Node n, binary b) where b contains a compressed text, performs the following actions: b is decompressed, stored into a container, and then the iteration AA_nextIter(Node n) is started, this iteration returns the next text in the container
- bool AA_hasReceivedText(Node n) returns true iff the text for n has been received

D.Initialization

SN restores the skeleton tree T_D and then the annotated tree $T_{A,D}$ (but it does not decompress text containers), then it sends the *skeleton* tree to RN:

SN: send(T_D)
RN: receive(T_D)

After the initialization, RN runs the following procedure:

```
SN_dfs(AnnotationTreeNode f) {
   for (every child c of f)
      if(isText(c))
         send(c, text of c);
      else {
         send(ANN(c));
         SN_dfs(c);
      }
} // SN_dfs()
```

For the RN, we show the recursive version:

```
RN_restore_recursive(SkeletonTreeNode f) {
   for (every child c of f)  //from left to right siblings
      if (is_Text(c)) {
         if (!AA_hasReceivedText(c))
            AA_getText(c);
         output(AA_nextTextIter(c));
         return;
      } else {
         if (!AA_isInit(c)) {
            receive(ann);
            AA_init(c);
            AA_store(c,ann);
         }
         while (AA_getFirst(c) > 0) {
            output("<" + tag(c) + ">");
            a(c)+=-1;
            RN_restore_recursive(c);
            output("</" + tag(c) + ">");
         }
         AA_delete(c);
   }
} // RN_restore()
```

For the XML file from Fig. 1 (a), in Fig. 3 we show packets that will be sent by SN_send() and the state of the annotated tree after each packet has been processed by RN_restore(), (the current node is bold, un-annotated nodes have annotation [1]). Note the last state (in the right bottom corner) shows the same annotated tree as in Fig 1 (b).

E. Querying Strategy

For queries formulated using a subset of XPath, the network node receiving a compressed data can query this data as it is being processed. Specifically, XSAQCT decompresses the skeleton tree, and annotations, and then decorates the tree with annotations. Depending on a type of the query, it can be immediately answered (for exact-match queries involving only tag names, e.g. /a/b/) or (for example to find the location of some text data) XSAQCT finds the location of the data, decompresses the appropriate data container, and completes the evaluation of the query. This type of lazy decompression makes the evaluation of queries more efficient.

IV. EXAMPLES OF APPLICATIONS

To consider possible applications of online XSAQCT, see Fig. 2, consider the network node N1, which produces XML data, to be sent to the network node N2, where they are compressed online by XSAQCT. N2 stores compressed data, which can be queried by the network node N3 sending queries. They can also be decompressed online by XSAQCT and sent to the network node N4. This node can either store uncompressed data, or they can be piped into any WWW application. Therefore, this figure shows the general architecture of our system. For example, for online decompression, input data may be piped into the compressor and sent over the Internet. On the receiving end, the data may be piped into two programs; one that collects the entire compressed document, and a second program which performs on-the-fly decompression. The decompressed data can be piped into any WWW application, such as a SOAP processor. The complete compressed data can be stored, and queried without having to decompress it.

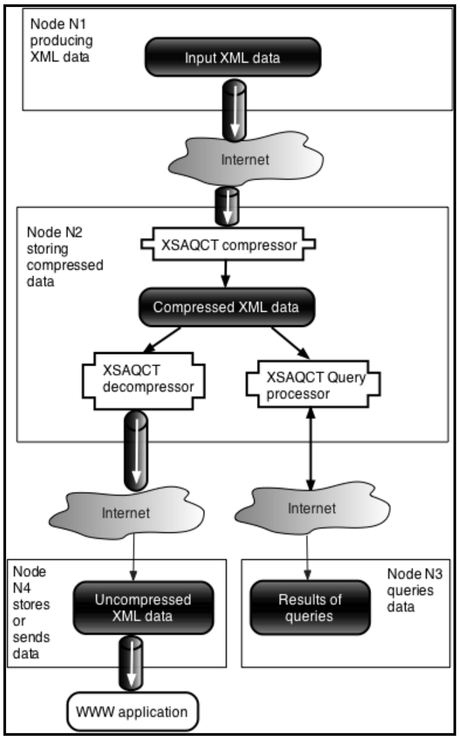


Fig. 2. Applications

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented an online XML compressor/decompressor XSAQCT suited for efficient implementation of online communication. We provided the brief outline of the implementation and results of the implementation.

In this version we did not attempt to handle XML mixed content or cycles, e.g., nodes with the consecutive children b, c and b. In the future version, we will remove these limitations. In addition, the future version will add more querying and updating facilities. Finally, we will add parallelization to the online compressor, based on [13]. To evaluate the effectiveness of online XSAQCT; specifically its compression and decompression and compression ratios, we will use three files of varying sizes: shakespeare.xml, dblp.xml and 1gig.xml. The first two files are taken from the Wratislavia corpus [18], while the last file is a randomly generated XML file, using xmlgen [19]. We will test our code by recording: (a) time to send a single uncompressed XML file D over the network from node N1 to node N2 and then compressing offline in N2, and (b) time to compress D online while sending from N1 to N2. Similarly, we will record (a) time to send a single compressed XML file D over the network from node N1 to node N2 and then decompressing offline.

Example**.**

For the XML file from Fig. 1 (a), in Fig. 3 we show packets that will be sent by SN_send() and the state of the annotated tree after each packet has been processed by RN_restore(), (the current node is bold, un-annotated nodes have annotation [1]). Note the last state (in the right bottom corner) shows the same annotated tree as in Fig 1 (b).

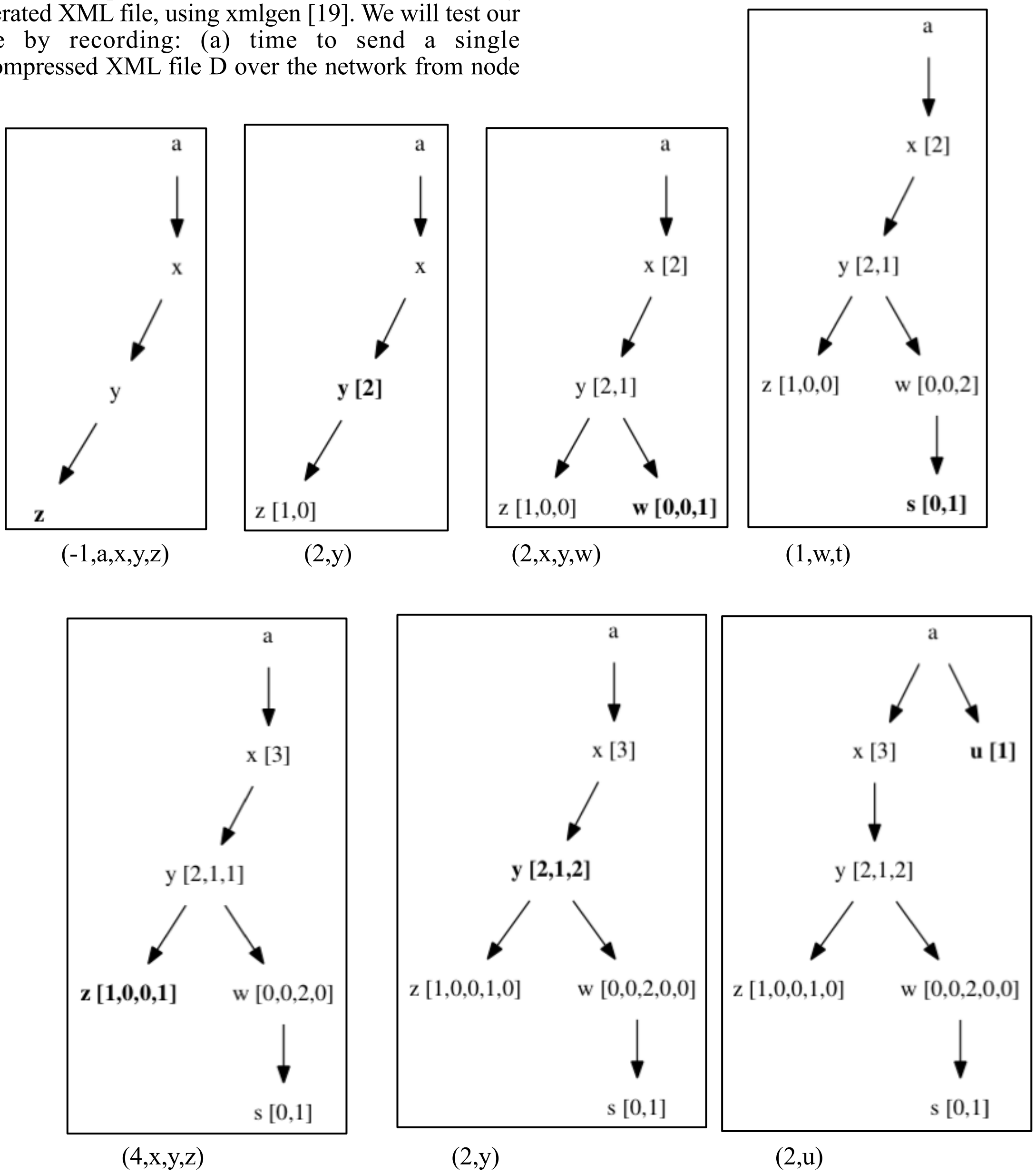


Fig.3 The state of the annotated tree after sending each packet

References

[1] W3C, *Extensible Markup Language (XML) 1.0 (Fifth Edition)*. http://www.w3.org/TR/REC-xml/, 2011. Retrieved on January 20, 2012.

[2] H. Liefke and D. Suciu, “XMill: an efficient compressor for XML data,” in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, 2000, pp. 153-164.

[3] P. Tolani and J. Haritsa, “XGRIND: a query-friendly XML compressor,” in *Proceedings of the 2002 International Conference. on Database Engineering*, 2002, pp. 225-34.

[4] A. Arion, A. Bonifati, G. Costa, S. D’Aguanno, I. Manolescu, and A. Pugliese, “XQueC: pushing queries to compressed XML data,” in *Proceedings of the 29th international conference on Very large data bases, 2010.*
Volume 29, Berlin, Germany, 2003, pp. 1065–1068.

[5] P. Skibiński and J. Swacha, “Combining efficient XML compression with query processing,” in *Advances in Databases and Information Systems*, 2007, pp. 330–342.

[6] Y. Lin, Y. Zhang, Q. Li, and J. Yang, “Supporting Efficient Query Processing on Compressed XML Files,” 2005.

[7] T. Müldner, C. Fry, J. K. Miziołek, and T. Corbin, “Updates of Compressed Dynamic XML Documents,” in *Eight International Network Conference*, 2010, pp. 315–324.

[8] I. Tatarinov, Z. G. Ives, A. Y. Halevy, and D. S. Weld, “Updating XML,” in *Proceedings of the 2001 ACM SIGMOD international conference on Management of data*, Santa Barbara, California, United States, 2001, pp. 413–424.

[9] S. Sakr, “An Experimental Investigation of XML Compression Tools,” *CoRR*, vol. abs/0806.0075, 2008.

[10] Müldner, T., Leighton, G., and Diamond, J., “Using XML Compression for WWW Communication,” presented at the IADIS International Conference WWW/Internet, 2005, pp. 459-466.

[11] T. Müldner, C. Fry, J. K. Miziołek, and S. Durno, “XSAQCT: XML Queryable Compressor,” Montréal, Canada, 2009.

[12] G. Leighton, T. Müldner, and J. Diamond, “TREECHOP: A Tree-based Query-able Compressor for XML,” *The Ninth Canadian Workshop on Information Theory*, Jun. 2005.

[13] T. Müldner, C. Fry, T. Corbin, and J. K. Miziolek, “Parallelization of an XML Data Compressor on Multi-cores,” presented at the PPAM, Torun, Poland, 2011.

[14] T. Müldner, J. K. Miziolek, and C. Fry, “Updateable Educational Applications based on Compressed XML Documents,” in *CSEDU (1)*, 2011, pp. 369-371.

[15] W3C, *Canonical XML*. http://www.w3.org/TR/xml-c14n, 2001. Retrieved on January 20, 2012.

[16] “Xerces,” *http://xerces.apache.org/xerces-j/*. [Online]. Available: http://xerces.apache.org/xerces-j/. Retrieved on January 20, 2012.

[17] *The gzip home page*. http://www.gzip.org/. Retrieved on January 20, 2012.

[18] *Wratislavia XML Corpus*. http://www.ii.uni.wroc.pl/ \textasciitildeinikep/research/Wratislavia/. Retrieved on January 20, 2012.

[19] *xmlgen - The Benchmark Data Generator*. http://www.xml-benchmark.org/generator.html. Retrieved on January 20, 2012.

[20] T. Müldner, G. Leighton, and J. Diamond, “Using xml compression for www communication,” in *Proceedings of the International Association for Developement of the Information Society (IADIS) International Conference WWW/Internet 2005 (ICWI 2005)*, 2005, pp. 459–466.

Physical Theories of the Evolution of Online Social Networks: A Discussion Impulse

Lutz Poessneck, Henning Hofmann, Ricardo Buettner
FOM Hochschule fuer Oekonomie & Management, University of Applied Sciences
Chair of Information Systems, Organizational Behavior and Human Resource Management
Arnulfstrasse 30, 80335 Munich, Germany
lutz.poessneck@gmx.de, henning.hofmann@gmx.de, ricardo.buettner@fom.de

***Abstract*—The evolution of online social networks (OSN) is a hot topic in computer science. Surprisingly a lot of research on this topic was also done by physicists – following the long history of studying networks in physics. To give an interdisciplinary discussion impulse, this paper intends to delineate the major physical theories of the evolution of (O)SN. Furthermore the paper presents a future research gap, which consists in the lack of adoption of the physical theory of preferential attachment on OSN by computer scientists.**

***Keywords*-online social networks; evolution; physical theories; interdisciplinary perspective.**

I. Introduction

In the last years, there has been a lot of research activities on OSN in different disciplines (e.g., computer science [1], [2], economics [3], sociology [4], and psychology [5], [6]). Understanding the structure of OSN, as well as the processes that shape them, is regarded as important [7]. "It would be useful to have efficient algorithms to infer the actual degree of shared interest between two users, or the reliability of a user (as perceived by other users). With respect to security, it is important to under- stand the robustness of such networks to deliberate attempts of manipulation [7, p. 31]."

One of the main research aims concerns the evolution of OSN. Despite of the rich spectrum of the mentioned disciplines there are huge differences in saturation on OSN evolution research. Therefore, three databases were queried on 19th August and 20th August 2011: ACM Digital Library, IEEE Xplore Digital Library and ISI Web of Knowledge. Search terms were 'evolution social networks', 'social network theory internet', 'social networks evolution internet' and 'evolution social network internet'. The selection criterion was 'most cited' on IEEE Xplore Digital Library, 'times cited' on ISI Web of Knowlegde and 'citation count' on ACM Digital Library. Searched were in abstract, title and content of articles. To complete these results, the databases ScienceDirect and SpringerLink were also queried on 3rd September 2011 for 'relevant' articles. Between all the findings were selected the 15 most relevant articles. The relevance was evaluated with regard to the abstracts. An article, that had been comparatively less quoted, was preferred under certain circumstances to a more often cited article, because of its ability to answer the research question.

Table I
15 most relevant articles

TITLE
A comparative study of social network models: network evolution models and nodal attribute models [8]
Emergence of a small world from local interactions [9]
Empirical analysis of an evolving social network [10]
Evolution of a large online social network [11]
Evolution of the social network of scientific collaborations [12]
Measurement and analysis of online social networks [7]
Microscopic evolution of social networks [13]
MySpace and Facebook: applying the uses and gratifications theory to exploring friend-networking sites [14]
Properties of on-line social systems [15]
Self-similar community structure in a network of human interactions [16]
Social Networking [17]
Structure and evolution of online social networks [18]
Structure and time evolution of an Internet dating community [19]
The Evolution of Social and Economic Networks [20]
The structure and function of complex networks [21]

The research revealed a great interest of the physicist community: six of the most relevant 15 articles were published in physical journals, including the most relevant article. Further articles came from Computer Science (4), Economic Theory (one), Mathematics (one), Psychology (one), Science (one), and Social Networks (one). For control purposes the database ISI Web of Knowledge was again queried on 19th November 2011. The queried terms were 'evolution social network internet', the selection criterion was 'most cited', the subject of the research where the abstract, the title and the content of articles. The research also showed the great interest of physicists. Between the first 25 results came eight from physical journals. Here the list of the further ranking of sciences: Computer Science (six), Health (two), Law (two), Biology (one), Management Science (one), Marketing (one), Psychology (one), Science (one), Social Networks (one), and Telecommunications Policy (one). As a result of these inquiries, the research question was redefined in: How physical theories explain the evolution of OSN? This paper focuses on the OSN, seen through the glasses of physics.

Why physicists are interested in OSN at all? "The study of networks has had a long history in mathematics and the sciences... [22, p. 1]"– but the recent time brought a renewal: the advent of modern database technology, which can process even huge amounts of data [15, p. 107]. "Being far larger than the datasets of traditional social network analysis, these networks are more amenable to the kinds of statistical techniques with which physicists and mathematicians are familiar [22, p. 6]". At the end of the 1990s the investigation of massive amounts of data with mathematical and physical techniques marked the beginning of a "new science of networks [22, p. 4]" (leading theorists: Albert-László Barabási, Mark Buchanan, Duncan J. Watts and Mark Newman [23, p. 57]). According to Barabási, Newman and Watts the new science distinguishes itself from the previous work in three ways: first by focusing on the properties of real-world networks, second by looking networks as evolving structures and thirdly by considering networks as dynamical systems [22, p. 4].

A. Method and structure of the article

The literature-based work systematically investigated the most substantial relevant databases (ACM Digital Library, IEEE Xplore Digital Library, ISI Web of Knowledge, ScienceDirect, SpringerLink) with regard to the physical theories of the evolution of networks. The 15 most relevant articles present research findings referring on physical theories of the evolution of networks – but these articles do not explain these theories in detail. Therefore a mere presentation of the research results would be incomprehensible for a reader who is not a network theorist. That is why we structure the research findings on the basis of the article "Scale-Free Networks" by the physicists Albert-László Barabási and Eric Bonabeau [24] who provide an introduction to current physical theories of the evolution of networks. This article is used as a framework for the general explanation of certain topics, which are then refined by the findings in the 15 most relevant articles.

The paper is organized as follows. Section II introduces the relevant terms. Section III discusses physical theories of the evolution of networks and their application on OSN. Section IV lectures criticism on the the physical theories about OSN. Section V explores how the physical theories of the evolution of OSN are adopted by computer scientists. Section VI shows limitations of this paper, presents a future research gap and open issues.

II. Clarification of terms

Evolution is considered as "the development or growth, according to its inherent tendencies, of anything that may be compared to a living organism (*e.g.,* of a political constitution, science, language, etc.); sometimes, contrasted with *revolution*. Also, the rise or origination of anything by natural development, as distinguished from its production by a specific act; 'growing' as opposed to 'being made' [25, p. 477]." A network is "a set of items, which we will call vertices or sometimes nodes, with connections between them, called edges... [21, p. 168]." A social network is a "set of people or groups of people with some pattern of contacts or interactions between them... [21, p. 172]." The term OSN is used in the sense of a social-networking site, defined by boyd and Ellison: "We define social network sites as web-based services that allow individuals to (1) construct a public or semi-public profile within a bounded system, (2) articulate a list of other users with whom they share a connection, and (3) view and traverse their list of connections and those made by others within the system. The nature and nomenclature of these connections may vary from site to site [1, p. 211]." Physical terms are explained in Section III, directly when physical theories are presented. The term OSN is always used in plural. The terms vertex and node are congruent.

III. Physical Theories on OSN

Among the 15 most relevant articles are [7], [11], [13], [15], [18] and [19] of particular interest: they are empirical studies that apply current physical theories of the evolution of networks on OSN – in the sense of definition of OSN given in Section II. Covered OSN are the Swedish community Pussokram.com [19], the Polish community Grono.net [15], the Community of the Polish Massive Multiplayer Online Role Playing Game Allseron.com [15], LastFM [15], the books admirer community Shelfari [15], Flickr [7], [13] and [18], YouTube [7], LiveJournal [7], Orkut [7], Yahoo 360° [18], Delicious [13], Yahoo Answers [13], LinkedIn [13] and the Chinese business community Wealink [11]. Other research subjects were databases of journals [12], email communications within universities [10] and [16] and communications via XFire, an instant messaging service for gamers [15].

A. Power-law distribution of vertex size

Barabási et al. experimented in 1998 with software to map how Internet sites are connected [24, p. 62]. The sites were sorted according to their size (the number of their links). Barabási and Bonabeau [24] expected to find a Poisson distribution of sizes: that sizes cluster around a mean value and sites with much more or fewer links are likely to be an exception. But the measurements refute the expectation: More than 80 percent of sites had less than four links, but in a minority of less than 0.01 percent had each site more than a thousand links. According to these results the World Wide Web is held together by very few, very large connection-rich vertices. The sorting of the sites with regard to their size revealed a "power law": the probability that any site has exactly k links is roughly proportional to $1/k^n$. The associated distribution curve does not have the pronounced peak at the typical size, but "is instead described by a continuously decreasing function [24, p. 63]."

Among the 15 most relevant articles were found that distribution of node size – the number of their incoming and outgoing connections – on Allseron.com, Grono.net, LastFM and Shelfari follows the power-law scaling form [15, p. 107]. Also Flickr, LiveJournal, Orkut and YouTube "show properties consistent with power-law networks [7, p. 36]." The distribution of vertices according to their size of Flickr and Yahoo 360° [18, p. 613] and Wealink [11, p. 1107] revealed a power law. By focusing on the microscopic vertex behavior of Flickr, Delicious, LinkedIn and Yahoo Answers it was also analytically shown that the edge initiation process can be captured by exponential vertex lifetimes and a "gap model" based on a power law [13, p. 470]. The degree distribution of vertices on Pussokram.com does not "fits a power-law form across the whole range observed [19, p. 165]." The degree is the number of edges connected to a vertex. However, the authors refer to a study of the French OSN nioki.com that has reported a power-law of the cumulative degree distribution. They conclude that "a closer inspection of our graphs… reveals a striking similarity in the functional form of the distribution. We therefore conclude that the dynamics shaping the degree distribution is to a large extent the same for the two communities [19, p. 165]."

B. Scale-freedom

The term scale-free means: there is no vertex size, which could be considered as "normal" and thus could apply as a measure [24, p. 62]. "Over the past few years, investigators from a variety of fields have discovered that many networks – from the World Wide Web to a cell's metabolic system to actors in Hollywood – are dominated by a relatively small number of nodes that are connected to many other sites. Networks containing such important nodes, or hubs, tend to be what we call "scale-free," in the sense that some hubs have a seemingly unlimited number of links and no node is typical of the others [24, p. 62]." Among the 15 most relevant articles it was found that the OSN Flickr, LiveJournal, Orkut and YouTube show scale-free properties [7, p. 32]. The authors of [7] calculated a value called 'scale-free metrics' that stands between 0 and 1 and measures the extent to which the graph of an OSN has a hub-like core. "A high scale-free metric means that high-degree nodes tend to connect to other high-degree nodes, while a low scale-free metric means that high-degree nodes tend to connect to low-degree nodes [7, p. 38]." The values are 0.49 for Flickr, 0.34 for LiveJournal, 0.36 for Orkut and 0.19 for YouTube. "All of the networks with the exception of YouTube, indicating that high-degree nodes tend to connect to other high-degree nodes, and low-degree nodes tend to connect to low-degree nodes [7, p. 38]."

C. Preferential attachment

Barabási and Bonabeau [24] attribute scale-freedom to two causes. First the older a vertex is, the more opportunities it had to build links. Therefore, vertices tend to be greater the longer they have been in the network. The second cause was called "preferential attachment". New vertices are connected preferentially with the major vertices and therefore major vertices are getting greater and greater over time. "… as new nodes appear, they tend to connect to the more connected sites, and these popular locations thus acquire more links over time than their less connected neighbors. And this "rich get richer" process will generally favor the early nodes, which are more likely to eventually become hubs [24, p. 65]."

Among the 15 most relevant articles the authors of [13] aim "to quantify the amount of preferential attachment that occurs in networks [13, p. 470]." For Flickr, Delicious, LinkedIn and Yahoo Answers they found that preferential attachment "is a reasonable model for edge destination selection [13, p. 465]." Using the statistical method of maximum-likelihood estimation, they show distortions in two assumptions of the preferential-attachment-theory: edge attachment by degree of vertices and edge attachment by the age of a vertex.

D. Small World

According to Watts and Strogatz [26] the connection topology of networks is neither completely regular nor completely random. "But many biological, technological and social networks lie somewhere between these two extremes [26, p. 440]." They are regular networks with increasing amounts of disorders. "We found that these systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. We call them 'small-world' networks, by analogy with the small-world phenomenon… [26, p. 440]." Watts and Strogatz refer to experiments of the social psychologist Stanley Milgram [27] in the 1960s, in which letters passed from person to person were able to reach an individual target in six steps. This was attributed to a few people who have had a lot of connections to other people ("hubs" in modern words) and have been addressed for transmission.

According to Barabási and Bonabeau [24] scale-free networks have also small-world properties. Even a large network with purely randomly placed connections has usually this property [24, p. 68]. If a person has one hundred acquaintances and any of them has again one hundred acquaintances, then 10.000 people are only two handshakes away from this person. And a million people are about three handshakes away [24, p. 68].

Studies have shown that the World Wide Web, scientific collaboration on research papers and general social networks have small-world properties [7, p. 32]. Among the 15 most relevant articles small-world properties were found for the OSN Allseron.com [15], Grono.net [15], LastFm [15], Shelfari [15], XFire [15], Flickr [7], YouTube [7], LiveJournal [7] and Orkut [7].

E. Clustering

According to Barabási and Bonabeau [24] the calculation in Section III/D has a hook: It is assumed that the acquaintances do not know each other. “In reality, there is much overlap [24, p. 68].” In fact, already in the second stage fewer than 10.000 people come together. This is because the society “is fragmented into clusters of individuals having similar characteristics (such as income or interests)... [24, p. 68].” Clustering is found in various networks. “At first glance, isolated clusters of highly interconnected nodes appear to run counter to the topology of scale-free networks, in which a number of hubs radiate throughout the system, linking everything. Recently, however, we have shown that the two properties are compatible: a network can be both highly clustered and scale-free when small, tightly interlinked clusters of nodes are connected into larger, less cohesive groups... [24, p. 68].”

Among the 15 most relevant articles the ‘clustering coefficient’ is introduced as a measure for the cluster. “The *clustering coefficient* of a node with *N* neighbors is defined as the number of directed links that exist between the node’s *N* neighbors, divided by the number of possible directed links that could exist between the node’s neighbors $(N(N-1))$ [7, p. 39].” Observed clustering coefficients are 0.313 for Flickr, 0.330 for LiveJournal, 0.171 for Orkut and 0.136 for YouTube [7, p. 39]. “The clustering coefficients of social networks are between three and five orders of magnitude larger than their corresponding random graphs. This unusually high clustering coefficient suggests the presence of strong local clustering, and has a natural explanation in social networks: people tend to be introduced to other people via mutual friends, increasing the probability that two friends of a single user are also friends [7, p. 39].”

IV. Criticism on physical theories on OSN

Criticism comes from the sociology. The sociologist Scott states: “Because of an apparent decline in the number of soluble theoretical problems that are left to resolve in their own discipline, a growing number of theoretical physicists have begun to explore the implications of some of their mathematical ideas for the explanation of social and economic phenomena [23, p. 55].” According to Scott, some physicians present their arguments as new - but they are not new at all. “The much-trumpeted innovations that lie at the heart of their ’revolution’ - the power law and hubs - are well-known and well-established findings of social network analysts [23, p. 62].” According to the sociologist “it is certainly the case that the terminology of scale-free distribution or power law was not used, but standard frequency distribution tables were used precisely in order to display this pattern [23, p. 60].”

Despite the criticism Scott attests the “new social physics” some good ideas and calls for an interdisciplinary exchange. “Much research in social network analysis has been static and cross-sectional ... This perspective has converged with uses of complexity theory and agent-based computational methods to begin to produce more powerful and productive examinations of longitudinal change [23, p. 64].”

V. Adoption of physical theories on OSN in computer science

This section investigates how physical theories on OSN are picked up by computer scientists. To measure this ACM Digital Library, IEEE Xplore Digital Library and ISI Web of Knowlegde were considered as the most relevant article collections regarding computer science and queried on 10th January 2012. Selection criterion was ‘most cited’ on IEEE Xplore Digital Library, ‘times cited’ on ISI Web of Knowlegde and ‘citation count’ on ACM Digital Library. Searched was in abstract, title and content of articles. Among the results of every searched phrase three articles were randomly chosen and proved that the terms used in the articles were terms in the appropriate sense. The discussion of the results follows in Section V.

Power-law distribution of vertex size: Search terms were “power-law distribution” + “online social network”. On ACM Digital Library 912 results were displayed. IEEE Xplore Digital Library showed 3 results, ISI Web of Knowlegde 2. Total number of findings: 964.

Scale-freedom: Search terms were “scale-free” + “online social network”. ACM Digital Library displayed 635 results. On IEEE Xplore Digital Library 37 results were shown. ISI Web of Knowlegde revealed 10 results. Total number of findings: 682.

Preferential attachment: Search terms were “preferential attachment” + “online social network”. On ACM Digital Library 316 results were displayed. IEEE Xplore Digital Library showed 40 results, ISI Web of Knowlegde 5. Total number of findings: 361.

Small World: Search terms were “small world” + “online social network”. ACM Digital Library displayed 14.874 results. On IEEE Xplore Digital Library results 784 results were shown. ISI Web of Knowlegde revealed 21 results. Total number of findings: 15.679.

Clustering: Search terms were “cluster” + “online social network”. On ACM Digital Library 6.393 results were displayed. IEEE Xplore Digital Library showed 3.614, ISI Web of Knowlegde 57 results. Total number of findings: 10.064.

VI. Conclusion

In this paper, we studied OSN, seen through the glasses of the physical approach of a "new science of networks". Some physical theories of the evolution of OSN are widely discussed among computer scientists: small world effect, clustering, power-law distribution of vertex size and scale-freedom. On the contrary, only very few publications of computer scientists deal with preferential attachment on OSN. Hence this topic was identified as a future research gap.

A. Future research gap: Preferential attachment on OSN

Physical theories on OSN are already picked up by computer science – but to varying degrees. Most discussed is the small world effect on OSN (15.679 results). Even the phenomenon of clustering on OSN (10.064 results) and the power-law distribution of vertex size in connection with OSN (964 results) attracts a lot of publications in computer science. Fewer articles deal with the scale-freedom of OSN (682 results). Surprisingly few articles have been published to the topic preferential attachment on OSN: 361 results. These are only 2.30 percent of the number of articles published to the small world effect on OSN – the number of published articles is out of proportion to the importance of the topic. According to Barabási and Bonabeau the process of preferential attachment occurs anywhere [24, p. 64]. "Likewise, the most relevant articles in the scientific literature stimulate even more researchers to read and cite them, a phenomenon that noted sociologist Robert K. Merton called the Matthew effect, after a passage in a Christian gospel: "For unto every one that hath shall be given, and he shall have abundance [24, p. 65]." In other areas – like the Internet and the U.S. biotech industry – preferential attachment has already been explored [24, p. 65]. Open questions, occurring through the identified research gap, are listed in Section V/C.

B. Limitations

A shortcoming of this paper is that the social networks, described in the 15 most relevant articles, are not homogeneous. Only a few are OSN in the sense of definition in Section II, for instance grono.net [15] or Wealink [11]. This applies to the database queries: During the calculation of the total number of results was not differentiated between OSN in the strict sense of definition in Section II and in a broader sense.

Another shortcoming lies in the lack of empirical studies among the 15 most relevant articles on the evolution of MySpace and Facebook, currently the largest OSN. To prove this, ACM Digital Library, IEEE Xplore Digital Library and ISI Web of Knowledge were queried on 26th January 2012. Search terms were 'Facebook', 'MySpace', 'social network' and 'evolution'. Selection criterion was 'most cited' on IEEE Xplore Digital Library, 'times cited' on ISI Web of Knowlegde and 'citation count' on ACM Digital Library. Searched was in abstract, title and content of articles. In the case of Facebook the query revealed 1261 results, in the case of MySpace 596 results. In both cases, the abstracts of the first 50 results of ACM Digital Library were checked, and even the abstracts of all results of IEEE Xplore Digital Library and ISI Web of Knowlegde. Articles that deal explicitly with the evolution of Facebook and MySpace were not found. There were, however, founded articles that already belong to the fundus of the 15 most relevant articles, for instance [7], [18] and [28].

According to Ellison et al. much of the existing academic research on Facebook has focused on identity presentation and privacy concerns [4]. Ryan und Xenos accentuate: "Despite the potential implications of Facebook use, there is a distinct lack of empirically derived theory in this area [6, p. 1658]." This could be, because Facebook is a relatively recent phenomenon, and as such, there has been limited opportunity for exploratory research [6].

C. Open Issues

The exploration of preferential attachment on OSN opens up a series of research questions for computer scientists. How is this process structured on OSN? How do the running of this process on OSN differ from the running in the offline world? How does the preferential attachment influence the dynamics of the evolution of OSN? What does preferential attachment for designing and conducting of OSN mean? How could a theory of preferential attachment be used to improve current OSN and to design new applications for OSN? How does preferential attachment influence the stability of OSN? Is it possible, to transfer the findings on preferential attachment, which have been obtained through the OSN, in the offline world?

Not only the "new science"-model, which was presented in this paper, uses preferential attachment, also other mechanisms do [28, p. 843]. "The transitive linking model..., which is based on continuously completing triangles with only an edge missing, is one such example [28, p. 843]." Another point of view is a fitness-based approach. "In any fitness-based approach, each node has its own fitness value and they are linked by the function of their fitness values [28, p. 843]." Hence further research could be done by computer scientists to compare and to integrate these different approaches and, if possible, to apply the integrated approach on OSN.

It could also stimulate research to include the perspectives of other sciences. OSN had attracted scientists of different backgrounds – at this point mostly physicists and computer scientists [11, p. 1110]. "However the main body in the virtual world is still persons in real world, thus as pointed out by Tim Berners-Lee – the "father of the World Wide Web", understanding the web community may also require insights from sociology and psychology every bit as much as from physics and computer science... [11, p. 1110]."

Acknowledgment

The authors gratefully acknowledge support from grant 17103X10 from the German federal ministry of education and research.

References

[1] D. M. Boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210–230, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1083-6101.2007.00393.x

[2] D. Richter, K. Riemer, and J. vom Brocke, "Internet social networking: Research state of the art and implications for enterprise 2.0," *Business & Information Systems Engineering*, vol. 3, no. 2, pp. 89–101, 2011. [Online]. Available: http://dx.doi.org/10.1007/s12599-011-0151-y

[3] S. P. Borgatti and D. S. Halgin, "On network theory," *Organization Science, Articles in Advance*, pp. 1–14, 2011, published online before print April 11, 2011.

[4] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of facebook "friends:" social capital and college students use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[5] Y. Amichai-Hamburger and G. Vinitzky, "Social network use and personality," *Computers in Human Behavior*, vol. 26, no. 6, pp. 1289–1295, November 2010.

[6] T. Ryan and S. Xenos, "Who uses Facebook? an investigation into the relationship between the Big Five, shyness, narcissism, loneliness, and Facebook usage," *Computers in Human Behavior*, vol. 27, no. 5, pp. 1658–1664, September 2011.

[7] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, ser. IMC '07. New York, NY, USA: ACM, 2007, pp. 29–42. [Online]. Available: http://doi.acm.org/10.1145/1298306.1298311

[8] R. Toivonen, L. Kovanen, M. Kivel, J.-P. Onnela, J. Saramki, K. Kaski, "Evolution of a large online social network: Network evolution models and nodal attribute models," *Social Networks*, vol. 31, pp. 240–254, 2009.

[9] J. Davidsen, H. Ebel, and S. Bornholdt, "Emergence of a small world from local interactions: Modeling acquaintance networks," *PHYS.REV.LETT.*, vol. 88, p. 128701, 2002.

[10] G. Kossinets and D. J. Watts, "Empirical analysis of an evolving social network," *Science*, vol. 311, pp. 88–90, 2006.

[11] H. Hu and X. Wang, "Evolution of a large online social network," *Physics Letters A*, pp. 1105–1110, 2009. [Online]. Available: http://dx.doi.org/10.1016/j.physleta.2009.02.004

[12] A. L. Barabási, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek, "Evolution of the social network of scientific collaborations," *Physica A*, vol. 311, pp. 590–614, 2002.

[13] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins, "Microscopic evolution of social networks," in *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2008, pp. 462–470. [Online]. Available: http://dx.doi.org/10.1145/1401890.1401948

[14] J. Raacke and J. Bonds-Raacke, "MySpace and Facebook: applying the uses and gratifications theory to exploring friend-networking sites." *Cyberpsychology and Behavior*, vol. 11, no. 2, pp. 169–174, 2008.

[15] A. Grabowski, N. Kruszewska, and R. Kosinski, "Properties of on-line social systems," *The European Physical Journal B*, vol. 66, pp. 107–113, 2008.

[16] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas, "Self-similar community structure in a network of human interactions," *Physical Review E*, vol. 68, no. 6, pp. 065 103+, December 2003.

[17] A. C. Weaver and B. B. Morrison, "Social Networking," *Computer*, vol. 41, no. 2, pp. 97–100, 2008.

[18] R. Kumar, J. Novak, and A. Tomkins, "Structure and evolution of online social networks," in *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2006, pp. 611–617. [Online]. Available: http://doi.acm.org/10.1145/1150402.1150476

[19] P. Holme, C. R. Edling, and F. Liljeros, "Structure and time evolution of an internet dating community," *Social Networks*, vol. 26, pp. 155–174, 2004.

[20] M. O. Jackson and A. Watts, "The Evolution of Social and Economic Networks," *Journal of Economic Theory*, vol. 106, no. 2, pp. 265–295, 2002.

[21] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

[22] M. E. J. Newman, A. L. Barabási, and D. J. Watts, *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[23] J. Scott and P. J. Carrington, *The Sage Handbook of Social Network Analysis*. SAGE Publications Ltd, 2011.

[24] A. L. Barabási and E. Bonabeau, "Scale-free networks," *Scientific American*, vol. 288, pp. 60–69, 2003.

[25] J. A. Simpson and E. S. C. Weiner, *The Oxford English Dictionary*. Oxford: Clarendon Press, 1989, vol. V.

[26] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, Jun. 1998. [Online]. Available: http://dx.doi.org/10.1038/30918

[27] S. Milgram, "The small-world problem," *Psychology Today*, vol. 1, no. 1, pp. 61–67, 1967.

[28] Y. Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 835–844.

Security is in the Eye of the Beholder: Security Perceptions and Challenges in Social Networks

Ioanna Dionysiou
Department of Computer Science
University of Nicosia
Nicosia, Cyprus
dionysiou.i@unic.ac.cy

***Abstract*—This paper discusses security as perceived by social networking participants. A conceptual security framework is presented that captures the security requirements that a user engaging in social networking activities may impose on other users, the social network provider, and a third-party user. We claim that even though the social network users seem to not value at the fullest extent the security that privacy that they are entitled, still the providers are responsible for supplying a secure infrastructure for user interactions that will protect users from security and privacy threats.**

***Keywords*-social networks; security; user perceptions**

I. Introduction

Even though social networking emerged as organized virtual communities in the last few years, its drastically-growing popularity is undisputed. Social networking sites such as Facebook, Linkedin, myspace, orkut, and twitter attract millions of users everyday. Social networking has been quickly adapted by the young population as the newest online trend, while there are very strong indications of a rapid growth amongst older users as well. According to a recent Nielsen report, *"social network and blogging sites are now the fourth most popular activity on the Internet"* [1], with the amount spent on these sites increasing by 63%. The popularity of social networks lies on the simple fact that they accommodate the exchange and sharing of information in an easy and intuitive manner for social, professional, and educational purposes. They even replace or supplement communications in the real world by diminishing barriers on physical location and time. Social networks provide opportunities to connect with friends, use short postings to inform friends on whereabouts, share videos and news, establish business contacts, advertise products, and campaign for various causes (political, social, etc.).

Social networks are subject to all common security vulnerabilities of the web [10] with their users being in even greater risk due to the implicit trust that governs these virtual communities. For example, users may show skepticism when receiving an email message that encourages them to click on a link or open an attachment, which is actually a malicious worm. However, they will click on such a link if it came from one of their social network connections. Needless to say, the sites that suffer more from security attacks are the most popular ones, and this realization has prompted several public and private bodies in lowering their tolerance of social networking activity during business hours. Besides the security concerns, privacy concerns also exist in social networks due to the vast amount of data that gets collected by the providers, allowing them to become digital *big brothers*. Personal and professional data could be exploited for a number of purposes, ranging from harming the system itself to increasing economic profits via data mining techniques. As an indicator of the monetary value of the stored data, the value of Facebook has been estimated to approximately $15 Billion.

Social networking represents the next generation of the Internet. It is here to stay. The aim of this paper is to investigate the security and privacy risks when interacting with social networking sites and present a security framework that those risks could be systematically assessed. Prior this discussion, a compact introduction to the structure and functionality of social networks is presented. Next, the findings of an empirical study that investigated the user perceptions of social network security is discussed. Security challenges of the construction of a global social network constitute the concluding part of this work.

II. Social Networking Sites (SNSs) Essentials

According to [8], social networking sites (SNSs) are web services that allow users to manage their profile within a bounded system, establish a list of connections, and finally traverse their connections' lists. However, this definition does not address the creation of new content and its dissemination among participants, which is after all the driving force behind social activities, either online or offline. Thus, a complete definition is one that relies on the *functional triangle* of social software that defines social software in terms of both information exchange and relationships. To be more precise, there are three primary functions of social software [3]:

1) Information management: creation, dissemination, and management of content, including searching

2) Self management: presentation of one's self to reflect various aspects of his/her personality
3) Relationship management: provision of profiles and management of connections

Hence, social networking sites are web services that support online social networks that provide to their members a platform that integrates a variety of information management and exchange tools (blogs, forums, instant messaging event management, media uploading applications, podcasts, etc.) as well as relationship management tools (profile construction, connection lists, searching). In addition, the SNS platform allows a user to express the aspects of him/herself that are considered to be important in the particular online community.

If we were to classify SNSs based on the type of information handled, then two categories arise: the first one is the group of SNSs that is used primarily for professional information dissemination, such as Linkedin that manages business contacts. The second group focuses on personal and private information and its character is more informal. Such an example is myspace.

A social network compromises of the SNS provider, the member users, and third-party sites that develop applications interacting with the SNS platform (e.g. in the case of Facebook). A user registers with the particular SNS and creates a profile by supplying basic, personal, contact and professional information, with an emphasis on the category that best represents the nature of social network. The user can use applications developed by the SNS providers or request use of a third-party application after getting authenticated by the SNS.

Facebook is selected from a plethora of social networks to serve as the example social network to demonstrate the functionality of a typical social networking site. The choice of Facebook is based on the undeniable fact that it is the largest and most feature-rich social network, with a rather broad set of privacy policies and thousands of third-party applications running on its platform. According to Jeff Rothschild, the Vice President of Technology at Facebook, there are currently 30000 servers supporting the operations of Facebook, with 25 terabytes of logging data managed daily on behalf of 300 million active users. Facebook develops its own in-house technologies to facilitate the sharing of information among its members, such as photos, notes, groups, events, posted items, video, marketplace, gifts. It supports features such as news feed, share, and wall for up-to-date info. The open Facebook API enables developers to integrate their own applications with Facebook and gain access to millions of users. However, the intriguing potentials of Facebook have an impact on the security and privacy of users, as it will be discussed later on.

III. Security and Privacy Risks

Security is in the eye of the beholder. The 2011 review of social networking sites as posted on the *www.toptenreviews.com* clearly suggests that the security of the most popular social networks ranges from *very good* to *excellent*. The evaluation criteria to assess the security of those sites were the following: support of privacy settings, block user feature, report spam feature, report abuse feature, and finally provision of safety tips.

This perception of security gives uninformed users a false reassurance. As a matter of fact, social networking sites suffer from a number of security vulnerabilities that could be exploited intentionally and accidentally [7], [24]. Facebook has suffered already XSS exploits, in the form of session hijacking and fake login pages. The infamous *harmless* Samy XSS worm shut down myspace in 2005 despite the fact that it only created inconvenience by adding the words *samy is my hero* to the top of every affected user page. Orkut users fell victims of a twitter-based scam, when they were lured to download a fake flash update that resulted in the launch of the worm that started harvesting google account details. Myspace and Facebook users were also the targets of the Koobface.a and koobface.b worms respectively. When a user of an infected machine log in to their social networking sites, fabricated messages were posted to the user's friends encouraging them to visit the malicious page.

Security and privacy in social networks as perceived by the users is also being investigated [4], [13], [20], [21]. Users seem to expect from the social network providers to support:

- Trustworthy environment: the community members should be able to trust each other, including applications.
- Privacy: users should be in control of privacy settings, which must be flexible and extensible
- Identity: even though the users are encouraged to reveal as little as possible to protect themselves from malicious acts, anonymity should be revoked when user harassment takers place
- Access control: users should have control over the content they generate by deciding its dissemination and revocation at any given time.
- Transparency: users must be informed how the collected data is used

Interestingly enough, there was no mention on vital security issues such as data integrity and confidentiality. The security and privacy problems do not only lie in the presence of design and implementation faults; the users carry their share of responsibility as well. If we were to examine the weakest links in the security of social networking, the investigation should have focused on all three participants groups as their actions have an impact on the overall security of the system:

- Users: The user behavior and user unawareness re-grading security are the primary factors that often lead to security and privacy problems. Bad habits also constitute a large fraction of the problems [10], [9]. User connectivity has become the primary objective of a significant number of users, who judge their importance by the numbers of friends they have. Experimental studies [18] have shown that almost half of the users agreed to accept as a friend someone they did not know, especially if a mutual friend existed between the requester and the target user [17]. This careless behavior increases the risk of being the victim of an attack, as a user with hundreds of friends is more likely to be subject to security breaches. In addition, users feel shielded from outside harm in online communities because they completely trust their connections. That's why they are more likely to click on a malicious link sent by a friend than if the link was sent via email, or they are willing to share personal information online than they would not normally do offline.

- SNS Provider: The SNS providers do not educate the users of risks of disclosing personal information [12]. For instance, users cannot control what their friends can reveal about them when using the tagging feature of Facebook. In addition, privacy tools and settings are not flexible or they are too complicated to be used properly by the average user. SNS providers do not provide the necessary security provisions for a number of security services, as it will be discussed below.
- Third-party: Social networks are complex systems that have their content and functionality enhanced by third-party applications. Rigorous methods are required to assess the security of the these system, and still is an open problem how to evaluate the security and safety of modules composition. As s result, malicious third-party applications could be launched via Facebook.

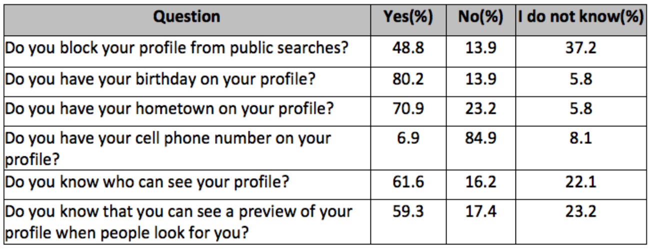
IV. User Perceptions of Security of Social Networks

In order to investigate the user perceptions of the security and privacy risks when interacting with social networks, a survey was conducted among Cypriot university students. The survey questionnaire (available upon request) focused on closed-ended questions that addressed factors involving most security services, such as authentication, confidentiality, integrity, access control, and privacy. It comprised of three sections. Part A collected demographic details, educational status, and internet usage information for the respondent. Part B aimed in gathering more information regarding the online activities a responded was involved in. Part C examined the perceptions that a social network user has on matters involving security risks, profile data disclosure, authentication process, privacy settings, privacy and confidentiality issues. At the end of the survey, the respondent was prompted to answer whether or not he/she will do anything different after taking the survey.

Questionnaires were collected during the period of October 2011 until December 2011, and the survey was conducted through personal interviews to assure the highest possible degree of accuracy for the received responses. The non-probability quota sampling method was employed with a sample of 109 users. The social network users were 86 and the non-users of social networks were 23. Starting with the findings for the first two parts of the survey, a total of 74% of the participants fell in the 18-34 age group, 86% of the respondents were listed as university students studied either in Cyprus or abroad, and 73% was using the internet on daily basis. Surprisingly, all social network users had a Facebook account, and approximately 10% also had a twitter account. It seems that Facebook is the dominant social networking site among Cypriot university students. When it comes to ways of accessing the social networking site, the most popular mean was using a laptop(45%), followed by a desktop (33%), and then a mobile phone (18%). The remaining users made use of tablets or another device.

The majority of the respondents claimed to be aware of social security risks in general (68.6%), however it is alarming that 15.1% is not aware of such risks and a percentage of 16.2% does not even know what a security risk is. As a follow up question, 32.5% responded positively when asked if they use a public machine to logon in a networking site and do not uncheck the *"keep me logged in"* button. Furthermore, 41.8% use the same password to log on to various social networking sites.

Figure 1 shows the response distribution for the questions referring to profile information and Figure 2 lists the responses for the profile settings. 6.9% of the users post their cell phone number on their public profile that is viewable at least by their connections and/or strangers. Approximately 40% of the respondents are not aware who can view their profile and are not concerned who has access to their information. A percentage of 36% is aware of the information that third-party applications collect, and a 27.9% even claims to know how the information is used and stored by such applications.

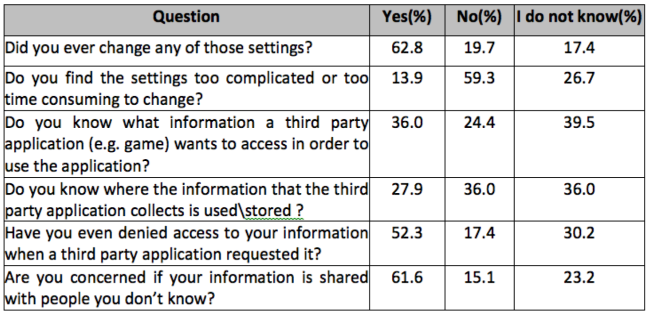


Question	Yes(%)	No(%)	I do not know(%)
Do you block your profile from public searches?	48.8	13.9	37.2
Do you have your birthday on your profile?	80.2	13.9	5.8
Do you have your hometown on your profile?	70.9	23.2	5.8
Do you have your cell phone number on your profile?	6.9	84.9	8.1
Do you know who can see your profile?	61.6	16.2	22.1
Do you know that you can see a preview of your profile when people look for you?	59.3	17.4	23.2

Figure 1. Response Distribution for Profile Question Set

Figure 3 shows the response distribution for questions

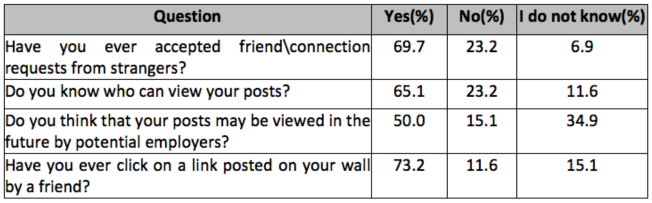
Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4



Question	Yes(%)	No(%)	I do not know(%)
Did you ever change any of those settings?	62.8	19.7	17.4
Do you find the settings too complicated or too time consuming to change?	13.9	59.3	26.7
Do you know what information a third party application (e.g. game) wants to access in order to use the application?	36.0	24.4	39.5
Do you know where the information that the third party application collects is used\stored ?	27.9	36.0	36.0
Have you even denied access to your information when a third party application requested it?	52.3	17.4	30.2
Are you concerned if your information is shared with people you don't know?	61.6	15.1	23.2

Figure 2. Response Distribution for Profile Settings Question Set

that involve a user's connections. An impressive 69.7% has accepted connection requests from strangers, showing that university students are willing to add into their circle users that they don't even know. Furthermore, 73.2% admitted that they click on a link posted by friends.



Question	Yes(%)	No(%)	I do not know(%)
Have you ever accepted friend\connection requests from strangers?	69.7	23.2	6.9
Do you know who can view your posts?	65.1	23.2	11.6
Do you think that your posts may be viewed in the future by potential employers?	50.0	15.1	34.9
Have you ever click on a link posted on your wall by a friend?	73.2	11.6	15.1

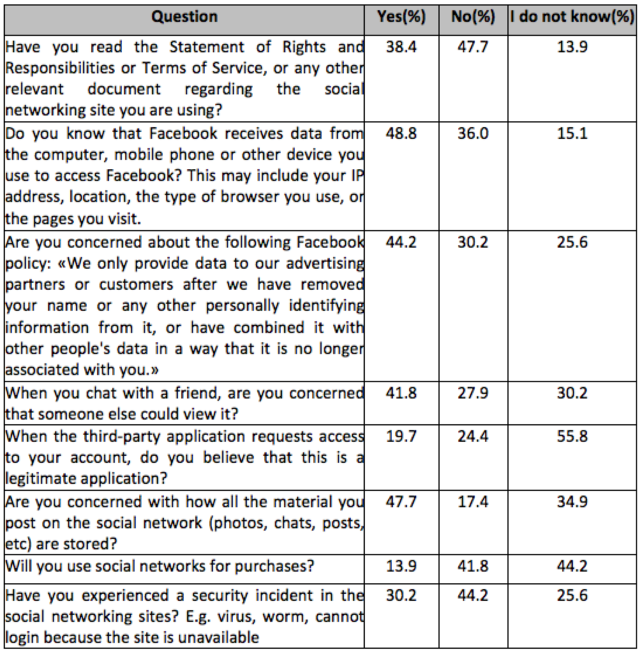
Figure 3. Response Distribution for Friends Question Set

Finally, Figure 4 reflects the replies of the respondents on privacy and other security risks. Less than half of the users have read the terms of service regarding the social networking site they are using. In addition, only half of them are aware of the information that the social network provider is collecting. Almost one fifth of the users believed that a third-party application is a legitimate application.

To conclude, it seems that not all users are concerned about privacy, access control of their information, storage or distribution of their personal data, confidentiality, and authentication. Besides, only 11.6% responded positively when asked if they will do anything different after taking the survey. This is an indication of lack of security-awareness among the target population, which is not always due to ignorance but it could be intentional as well.

V. Security Framework

Even though the social network users seem to not value at the fullest extent the security that privacy that they are entitled, still the providers are responsible for supplying a secure infrastructure for user interactions that will protect users from security and privacy threats. To assess and evaluate the security model of a social network, a systematic approach is needed to define the security requirements and characterize the approaches to satisfy them [23]. For our purposes, the security services required by a social networking site are the standard security services as defined by X.800: user authentication, data integrity, data confidentiality, data availability, and access control. Privacy, the ability to hide personal information from the system, is also a required service due to the vast volumes of data collected by both the provider and third-parties. Table I illustrates the comprehensive security and privacy framework for social networking, where services are established with connection to the system participants. In the discussion below, the focus is on the user-oriented requirements. The requirements imposed on the user by the SNS or the third party are outside the scope of this discussion.



Question	Yes(%)	No(%)	I do not know(%)
Have you read the Statement of Rights and Responsibilities or Terms of Service, or any other relevant document regarding the social networking site you are using?	38.4	47.7	13.9
Do you know that Facebook receives data from the computer, mobile phone or other device you use to access Facebook? This may include your IP address, location, the type of browser you use, or the pages you visit.	48.8	36.0	15.1
Are you concerned about the following Facebook policy: «We only provide data to our advertising partners or customers after we have removed your name or any other personally identifying information from it, or have combined it with other people's data in a way that it is no longer associated with you.»	44.2	30.2	25.6
When you chat with a friend, are you concerned that someone else could view it?	41.8	27.9	30.2
When the third-party application requests access to your account, do you believe that this is a legitimate application?	19.7	24.4	55.8
Are you concerned with how all the material you post on the social network (photos, chats, posts, etc) are stored?	47.7	17.4	34.9
Will you use social networks for purchases?	13.9	41.8	44.2
Have you experienced a security incident in the social networking sites? E.g. virus, worm, cannot login because the site is unavailable	30.2	44.2	25.6

Figure 4. Response Distribution for Security Risks Question Set

Table I
Security and Privacy Framework

	user-user	user-SNS provider	user-Third Party
authentication	no	yes	no
integrity	yes?	yes?	yes?
confidentiality	no?	no?	no?
availability	yes	yes	yes?
access control	yes?	no	yes?
privacy	yes?	no	yes?

A. Authentication

Authentication is one of the security services that is provided by almost all social networks. It refers to the assurance that the communicating entity (user, provider, third party) is the one that it claims to be. In order to implement the authentication service, credentials such as username (or email) and password need to be supplied by the unauthenticated user, and upon verification the user is either authorized to log on or access is not granted. In the case of Facebook, an SSL connection is established during the authentication phase so that the message exchange will be protected from eavesdropping. An authenticated user is presented with a session key that is used throughout the active session for any further authentication purposes.

When a user interacts with a third-party application, the authentication process will still be performed by the SNS provider. The third-party server does not perform any authentication on the user. Similarly, a user does not have the means to authenticate another user; there are no tools or mechanisms to verify the identity of another user. This is especially problematic when anonymity is viewed favorably by a number of users in order to protect their identity.

B. *Integrity*

Integrity refers to the assurance that the data has not been altered during its transmission to its indented destination. Due to the proprietary nature of the majority of social networks and the non-disclosure of technical specifications of the built-in or third-party applications, it is nontrivial to assess whether or not data integrity is part of the security model of the system and accompanied applications. There have been no incidents of message alteration (even though message fabrication has been witnessed), thus it could be assumed that some sort of message authenticator is generated that verifies the authenticity of the message. Taking Facebook as the example, the traffic among the external participants is digitally signed; however there is no description of how messages of built-in chatting applications are authenticated.

C. *Confidentiality*

When assessing the confidentiality strength of social systems, one needs to take into consideration the underlying purpose of these systems. The original goal was to facilitate various forms of communication among interacting parties. Secrecy was not a main concern, whereas access control and privacy were top priorities. But, with the increase of sophisticated attacks by knowledgeable hackers, confidentiality should also be of an equal concern. Currently, it is not clear how the network servers of the social networks interact with each other, and what security protocols are using.

Consider chatting applications. It is well-documented that the .Net Messenger Service allows unencrypted traffic, making the wiretapping of such conversation subtitle to both passive and active attacks. Facebook Chat was found to be subject to similar problems and has already started preparing a new interface which will be based on Jabber's XMPP (extensible messaging and presence protocol) that uses encryption to protect the secrecy of the communicated messages.

However, it may not be performance-wise to encrypt all traffic that goes through the social network. Trade-offs have to considered and perhaps the user could either opt-in or opt-out when it comes to encrypting communication sessions for different applications. Moreover, users could increase or increase the encryption strength, but with a monetary cost.

D. *Availability*

Availability is a system property where resources will be accessible and usable upon demand by an authorized system entity. Social networks suffer availability of service when denial of service attacks are launched due to either implementation vulnerabilities that get exploited or infected users that are used as points of launching worms and trojan viruses. Users expect their public profile information to be available to other users according to their preferences and they also anticipate that all features will be available whenever they want to use them. Users have the same availability demands from third-party applications as well – however, there are not any imposed availability requirements on the later applications. Needless to say, the more unavailable they are, the more users will abandon using their applications.

E. *Access Control and Privacy*

Social networks emphasize access control and privacy as the two most important pillars of their security model. Users have strong expectations for privacy on social networking sites and they believe that it is the responsibility of the SNS providers to protect personal and user-generated content.

The two terms are often used interchangeably as they are both associated with restricting access to user data. However, privacy involves more than controlling who can access what; it allows a user to be part of the environment without leaving any traces and enables his/her easy and permanent withdrawal without any evidence of the prior interactions. It can be claimed that the design of social networks partially implements both privacy and access control.

Starting with the user-to-user access control, social networks offer profile ”privacy”, meaning that the user configures privacy settings that explicitly specify the group of users that are granted access permission to various profile properties. This is a course-grained access control that handles a limited number of access groups such as friends and everybody. However, there is the option to block users. Once the data is accessible by others, the owner of the data has no control over its further dissemination and usage. As far as privacy is concerned, social networks such as linkedin and facebook support the search feature that control who can search for the user and the ways to get in contact.

Third-party applications are granted second-degree access permissions, resulting in gaining access not only to the data of the user who authorized the application but also getting access to friends' data. In a sense, applications become automatically friends of the user. The application developers are obliged, as dictated by the Terms of Service, to display a warning screen asking the user's consent in accessing data; this is quite meaningless as the user is given no choice to restrict access to information that the application does not need or provide anonymized data. Once the application is authorized by the user, social network providers have no way to check how the information is used by the third-party application; they only have the developers' consent that they will observe the Terms of Service.

And when it comes to the SNS provider, there are no technical obstacles to prevent access to all user data, supplied and generated, and further manage it as the provider sees appropriate. It is important to note that the users volunteer to abandon their rights to privacy by agreeing with the Terms of Service. For instance, Facebook explicitly specifies that personal information is stored and web site information (browser type, IP address) is stored from the user's browser using persistent cookies. In addition, according to the Facebook Terms of Service there is a wide range of information that Facebook gathers about a user*"...We receive data about you whenever you interact with Facebook, such as when you look at another person's profile, send someone a message, search for a friend or a Page, click on an ad, or purchase Facebook Credits...We receive data from the computer, mobile phone or other device you use to access Facebook. This may include your IP address, location, the type of browser you use, or the pages you visit...When we get your GPS location, we put it together with other location information we have about you (like your current city). But we only keep it until it is no longer useful to provide you services."*. In other words, whatever a user posts, views, searches, exchanges is stored on the Facebook servers.

VI. Social Network Challenges

The evolution of social networks into applications that span the web with millions of users plugged in offers new opportunities and challenges in the technological, economic, and social arenas. Below is a list (note: this list is by no means exhaustive) of security and privacy issues in each of these directions that are anticipated to be addressed in the near future.

A. Technological Directions: Global Social Ecosystem

One of the technological challenges in building a social ecosystem is how to achieve interoperability among SNSs. Blosser and Zhan [6] explain that in order to build a collaborative social network, three main issues have to be addressed, one of them being how to combine the data of the various social network providers while preserving user privacy and provider confidentiality. OpenSocial [16][15] is a framework that interlinks social networks that support its API. However, there is no mention on how security and privacy are implemented in this network of social networks.

The second challenge focuses on the sociological aspect of a global social network [19]. Aggregating audience of different communities implies the merging of multiple identities that users may have in those communities. However, the ability of a user to have different identities and portray the self to other in various ways will be simply disabled by the interconnection of social networks. There must be ways to protect the various roles and data of a user in this interconnected network: the professional role and the social role must be clearly distinguished as they are in real life.

B. Economic Directions

It has been observed that people tend to share the same interests with their friends, and this feature of *homophily* is vital if social networks were to be used for advertising. Various aspects of online advertising in social networks have been the subject of research works that present findings on how relevant online relationships are to advertising. The goal is to match an ad to a user. A recent study by Bagherjeiran and Parekh [5] investigated whether or not social network links are relevant to the targeted ads and how social information could be used in targeting methods to predict user response rates. It has been shown that the response rate on ads is indeed proportional to the number of connections who have responded in the past. They have hinted that relevant advertising will be more effective than viral spam.

The advertising business is already seeking ways to partner with social networks and gain access to the vast number of users that could be the target audience for their advertisements [2]. Mining social networks for viral marketing will be the future of advertising [22], with serious implications on the privacy of the user data.

C. Social Impacts

Web-based social networking is also transforming social habits, especially of the youth, by shifting from face-to-face communication to online interactions. It is argued that social networking fulfills a human need, that of gossiping. The largest the size of your friends group, the more efficient the dissemination of gossip becomes. However, is this an evolutionary shift that will change the way we operate or will it diminish as years go by?

VII. Conclusions

The popularity of social networking still exhibits an exponential growth, despite well-known and documented privacy and security breaches. The harm that a user may experience depends on how much the user engages in social networking activities. Social networks are complex systems and it is expected to observe security vulnerabilities from time to time.

However, could it be the case that we are reaching a new era where perhaps there is no such a thing as privacy anymore? The ability to collect data and monitor activities has serious impact to the users' privacy. Third-party companies could correlate public profile data and sell their finding to credit-card rating companies, insurance companies, employers, etc. That brings the question of what happens next. Shall users become more alert regarding the consequences of their interactions? Should a code of etiquette together with violation consequences [22] be established as part of the terms of service? Should security be transparent to the user [14] or security preferences will be specified and observed via service-level agreements for fine-tune security and privacy based on the interaction [11]?

References

[1] *Global Faces and Networked Places: A Nielsen report on Social Networkings New Global Footprint*. The Nielsen Company, March 2009.

[2] Many online social networks leak personal information to tracking sites, new study shows, August 2009.

[3] Richter A. and Koch M. Social software - status quo, 2007.

[4] Esma Aimeur, Sebastien Gambs, and Ai Ho. Upp: User privacy policy for social networking sites. In *Internet and Web Applications and Services, International Conference on*, pages 267–272. IEEE Computer Society, 2009.

[5] Abraham Bagherjeiran and Rajesh Parekh. Combining behavioral and social network data for online advertising. In *ICDMW '08: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*, pages 837–846. IEEE Computer Society, 2008.

[6] Gary Blosser and Justin Zhan. Privacy preserving collaborative social network. In *2008 International Conference on Information Security and Assurance*. IEEE Computer Society, 2008.

[7] Joseph Bonneau, Jonathan Anderson, and George Danezis. Prying data out of a social network. In *2009 Advances in Social Network Analysis and Mining*, pages 33–40. IEEE Computer Society, 2009.

[8] D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

[9] John Breslin and Stefan Decker. The future of social networks on the internet: The need for semantics. *IEEE Internet Computing*, 11(6):86–90, 2007.

[10] Steve Mansfield Devine. Anti-social networking: exploiting the trusting environment of web 2.0. *Network Security*, 2008(11):4–7, 2008.

[11] Ioanna Dionysiou, Dave Bakken, Carl Hauser, and Deborah Frincke. Formalizing end-to-end context-aware trust relationships in collaborative activities. In *International Conference on Security and Cryptography (SECRYPT08)*, pages 546–553, 2008.

[12] Ai Ho, Abdou Maiga, and Esma Aimeur. Privacy protection issues in social networking sites. In *ACS/IEEE International Conference on Computer Systems and Applications,*, pages 271–278. IEEE Computer Society, 2009.

[13] Amela Karahasanovic, Petter Bae Brandtzg, Jeroen Vanattenhoven, Bram Lievens, Karen Torben Nielsen, and Jo Pierson. Ensuring trust, privacy, and etiquette in web 2.0 applications. *Computer*, 42(6):42–49, 2009.

[14] Ryan Layfield, Bhavani Thuraisingham, Latifur Khan, Murat Kantarcioglu, and Jyothsna Rachapalli. Design and implementation of a secure social network system. In *IEEE Intelligence and Security Informatics 2009*, pages 236–247. IEEE, 2009.

[15] J. Mitchell-Wong, R. Kowalczyk, A. Roshelova, B. Joy, and H. Tsai. Opensocial: From social networks to social ecosystem. In *2007 Inaugural IEEE International Conference on Digital Ecosystems and Technologies*, pages 361–366. IEEE, 2007.

[16] Juliana Mitchell-Wong, Ryszard Kowalczyk, and Bao Quoc Vo. Social network profile and policy. In *2008 IEEE International Workshop on Policies for Distributed Systems and Networks*, pages 207–210, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[17] Frank Nagle and Lisa Singh. Can friends be trusted? exploring privacy in online social networks. In *2009 Advances in Social Network Analysis and Mining*, pages 312–315. IEEE Computer Society, 2009.

[18] Jan Nagy and Peter Pecho. Social networks security. In *Third International Conference on Emerging Security Information, Systems, and Technologies*, pages 321–325. IEEE Computer Society, 2009.

[19] Martin Pekarek and Stefanie Potzsch. A comparison of privacy issues in collaborative workspaces and social networks. *Identity in the Information Society*, pages 1–13, 2009.

[20] Cynthia Putnman and Beth Kolko. Getting online but still living offline: the complex relationship of technology adoption and in-person social networks. In *2009 Advances in Social Network Analysis and Mining*, pages 33–40. IEEE Computer Society, 2009.

[21] L. Sorensen and K.E. Skouby. Next generation social networks - elicitation of user requirements. In *IEEE 19th International Symposium on Personal, Indoor, and Mobile Radio Communications*, pages 1–5. IEEE, September 2008.

[22] Steffen Staab, Pedro Domingos, Peter Mika, Jennifer Golbeck, Li Ding, Tim Finin, Anupam Joshi, Andrzej Nowak, and Robin R. Vallacher. Social networks applied. *IEEE Intelligent Systems*, 20(1):80–93, 2009.

[23] William Stallings. *Network Security Essentials: Applications and Standards*. Pearson Prentice Hall, Upper Saddle River, NJ, USA, 3rd edition, 2006.

[24] M. Tubi, R. Puzis, and Y. Elovici. Deployment of dnids in social networks. In *2007 Intelligence and Security Informatics*, pages 59–65. IEEE, 2007.

Visual Storytelling with Snapshots applied to Official Statistics

Patrik Lundblad, Tobias Åström, Mikael Jern
[1]National Centre Visual Analytics, NCVA
Linköping University, Sweden
e-mail: patrik.lundblad@liu.se, tobias.astrom@liu.se, mikael.jern@liu.se

***Abstract*—The paper focuses on “visual storytelling” – exemplified through telling stories about official statistics development over time that could shape economic growth and well-being. Discoveries are made that draw the user into reflecting on how life is lived - and may be improved - from one region to another. In addition, the user can interactively participate in the web-based process, which is important to the education and dissemination of public statistics. A demonstrator World eXplorer with storytelling and an integrated snapshot mechanism is introduced, programmed in Adobe’s ActionScript and is based on the geovisual analytics paradigm. An interactive visual story mechanism assists the author to improve a reader’s visual knowledge through reflections on how life is lived by using a variety of demographics, such as healthcare, environment, and educational and economic indicators. Educators can develop interactive teaching material based on this storytelling mechanism and avoid boring statistics presentations. Integrated snapshots can be captured at any time during an explorative data analysis process and they become an important component of an analytical reasoning process. Students can access geovisual applications and explore statistical relations on their own guided by the stories prepared by the teachers. With the associated science of perception and cognition in relation to the use of multivariate spatio-temporal statistical data, this paper contributes to the growing interest in geovisual statistics analytics.**

Keywords-Visual Storytelling, Geovisual analytics, web-based learning, statistics visualization, statistics database

I. INTRODUCTION

The “participative web” is increasingly utilised by intelligent web services, which empower developers to customise web-enabled visualization applications that contribute to collaboration and communicate visual content (Figure 1). In this context, we introduce a collaborative geovisual analytics framework, based on the principles for Visual Analytics [1], for public statistics based on interactive statistics visualization [2] and its increasing role in sharing, collaborating and communicating gained knowledge. The collaboration and publication process facilitates storytelling aimed at producing statistical news content in support of an automatic authoring process. The author simply presses a button to publish the knowledge gained from a visual interactive discovery process. Storytelling, in our context, is about telling a story with associate snapshots about the statistics data and the related analytics reasoning processes on how knowledge was obtained. Storytelling within an interactive web context could more engagingly draw the user into reflections and sometimes change a perspective altogether. The story is placed in the hands of those who need it, e.g., policy and decision makers, teachers and also informed citizens. Stories enable a leap in understanding by the user in order to grasp how statistical indicators may influence society. The conceptual approach and framework of the geovisual analytics storytelling implementation is based around three complementary characteristics:

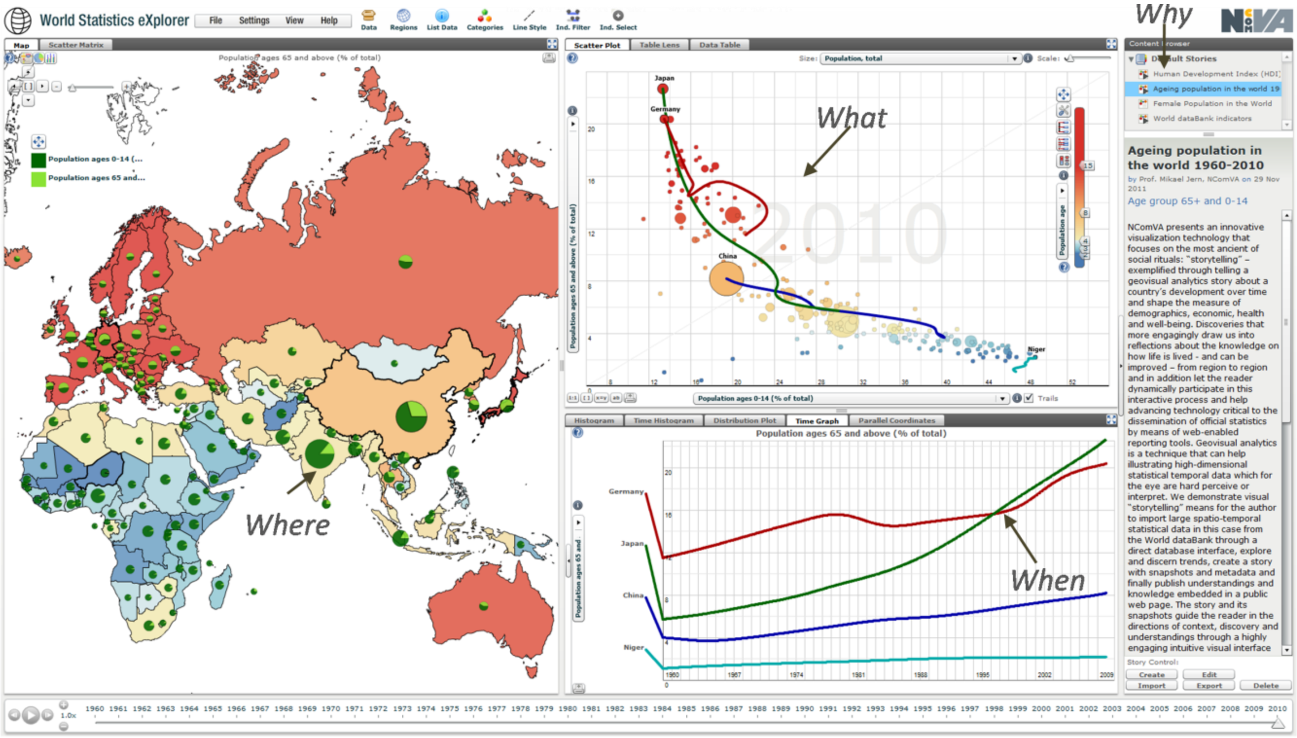
- **Authoring (eXplorer):** data provider and manager, several motion visual representations, coordinated views, map layers, analytic tools (dynamic query, filter, regional categorization, profiles, highlight), and dynamic colour scale.
- **Tell-a-story:** snapshot mechanism that captures an interactive scenario (active views and indicators, attributes, time step, regions), and metadata with hyperlinks.
- **Publisher (Vislet):** import stories and create HTML code for embeddable interactive motion visual representations for publishing on a web site, interactive report or blog.

The rest of this paper is structured as follows. In section II related work is presented followed by our work on storytelling in section III. The paper is concluded with section IV presenting our conclusions.

II. RELATED WORK

The importance to snapshot exploration sessions and then reuse them for presentation and evaluation within the same environment was demonstrated by MacEachren [3] and Jern [4] in geovisualization, and incorporated features to integrate them into electronic documents.

A variety of software is now available for creating snapshots. One of these is CCMaps [5], which is a conditioned choropleth mapping tool where the user can create snapshots of events and then reuse them for presentation. Another method is called “Re-Visualization” [6] and is used in the tool ReVise, which captures the analysis sessions and allows them to be reused. The Web-based Analysis and Visualization Environment (Weave) [7] is a Framework that uses session states and stores them on a server for later reuse. Another application that uses a similar approach where the user stores the data online is Many Eyes [8], which is a public website where novice users can upload their own data, create dynamic visualizations and participate in discussions. But, Many Eyes seems limited to showing



Ageing population in the world 1960-2010

by Prof. Mikael Jern, NComVA on 29 Nov 2011

Age group 65+ and 0-14

NComVA presents an innovative visualization technology that focuses on the most ancient of social rituals: "storytelling" – exemplified through telling a geovisual analytics story about a country's development over time and shape the measure of demographics, economic, health and well-being. Discoveries that more engagingly draw us into reflections about the knowledge on how life is lived - and can be improved – from region to region and in addition let the reader dynamically participate in this interactive process and help advancing technology critical to the dissemination of official statistics by means of web-enabled reporting tools. Geovisual analytics is a technique that can help illustrating high-dimensional statistical temporal data which for the eye are hard perceive or interpret. We demonstrate visual "storytelling" means for the author to import large spatio-temporal statistical data in this case from the World dataBank through a direct database interface, explore and discern trends, create a story with snapshots and metadata and finally publish understandings and knowledge embedded in a public web page. The story and its snapshots guide the reader in the directions of context, discovery and understandings through a highly engaging intuitive visual interface

Figure 1. World eXplorer [9] based on the geovisual analytics concept "Where-What-When-Why" based on 3 time-linked views showing a worldwide ageing population during 1960-2010; map (age 65+), scatter plot (age 65+ vs. age 0-14) and time chart (65+); comparing 4 countries Nigeria, China, Japan and Germany. The story is published to the right side and includes linked snapshots. Users learn that Japan followed by Germany represents the countries with highest elderly population, while Niger has has not changed. The users can interact and change indicators to discover reasons behind this trend.

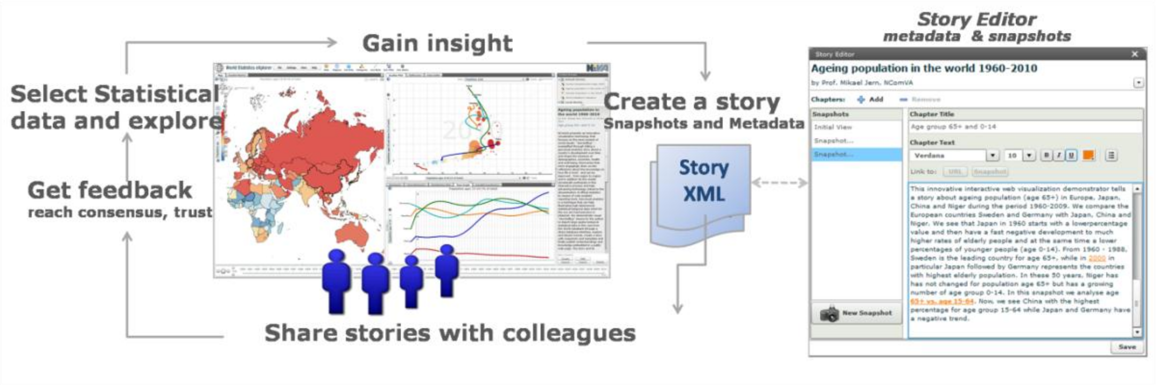
only one visualization at a time and has no animation facilities for time sequences. For many of these systems, the snapshot has to be loaded into the same application environment as the one that created it, which puts a restriction on usage and sharing if the application requires a software licence. Such applications may not be easily accessible to team members without installing external software [10]. In this context, we introduce a web compliant layered component toolkit with a snapshot mechanism that captures, re-uses and shares active properties for individual functional components. We have demonstrated [11] that such an implementation could provide a more open and collaborative geovisual analytics framework for public use. Collaborative geovisual analytics tools have been proved to work well with statistics data [12]. Initial tests have also shown significant potential [13, 14] when making them accessible to educators and their students. With the availability of current web enabled geovisual analytics tools it is appropriate to explore the possibilities of using these tools within schools and allow educators to use these tools in different application areas [15]. There is significant future potential for educators to present and explore scientific data sets together with students.

III. STORYTELLING

Storytelling within a participative web context, could more engagingly draw the user into exciting reflections and sometimes change a perspective altogether. The story is placed in the hands of the users as an interactive guided learning experience to achieve a more complete understanding of the knowledge using descriptive metadata integrated with "memorized interactive visualization views" or "snapshots" and external web links to relevant information. Visual storytelling is in this scenario an approach of telling more vital and engaging stories through interactive web-enabled visualizations. The snapshot mechanism helps the author of a story to highlight data views of particular interest and subsequently share or guide others to significant visual discoveries. These interactive events in an analytical reasoning process can be captured at any time during an explorative data analysis process and represent an important part of an analytical storytelling process (Figure 2).

A. *Snapshots*

When exploring and making sense of comprehensive statistics data there needs to be a way of storing discoveries in a coherent and cognitive workspace, which can be organized, navigated and maintained within the application so that discoveries may later be loaded. The GAV Flash toolkit [16, 17] includes such means by giving the analyst the option of capturing, saving and packaging the results of an exploration "gain insight" process. The insights are captured in a series of "snapshots" that could help the analyst to highlight views of particular interest and subsequently guide other analysts to follow important discoveries. Snapshots are created as a series of visualization captures during the exploration process and form the foundation for a story.



Story Editor

Ageing population in the world 1960-2010

by Prof. Mikael Jern, NComVA

Chapters: Add Remove

Snapshots: Initial View, Snapshot..., Snapshot...

Chapter Title: Age group 65+ and 0-14

Chapter Text: Verdana 10

Link to: URL Snapshot

This innovative interactive web visualization demonstrator tells a story about ageing population (age 65+) in Europe, Japan, China and Niger during the period 1960-2009. We compare the European countries Sweden and Germany with Japan, China and Niger. We see that Japan in 1960 starts with a lowerpercentage value and then have a fast negative development to much higher rates of elderly people and at the same time a lower percentages of younger people (age 0-14). From 1960 - 1988, Sweden is the leading country for age 65+, while in 2000 in particular Japan followed by Germany represents the countries with highest elderly population. In these 50 years, Niger has has not changed for population age 65+ but has a growing number of age group 0-14. In this snapshot we analyse age 65+ vs. age 15-64. Now, we see China with the highest percentage for age group 15-64 while Japan and Germany have a negative trend.

New Snapshot

Save

Figure 2. The analyst is the author in this storytelling loop

In a typical scenario the author selects relevant attributes for a snapshot, e.g., time step, highlighted regions for comparisons, class values for colour legend, filters inquiry conditions for a reduced dataset and finally highlights the "discoveries" in the world map view from a certain angle. When the analyst presses the capture button all the components within the application stores their properties, thus creating a state, which can later on be recreated (Figure 3). The snapshot is then inserted in the story by creating a hyperlink in the text that is highlighted. When the reader follows the story the reader can click the highlighted link and the state of the visualization when captured is recreated so the reader has the same view as the author. The reader may then also make changes to the application such as changing filter values or highlighting other areas of interest. This new discovery may then be stored as a new snapshot, which can be inserted into the story or replace the old snapshot.

In our story editor (Figure 6) the author has access to all stored snapshots in a list on the left side and can easily go to one of them and change the parameters. The author can also format the text and insert url links to other resources on the web. A story may also contain multiple chapters so that multiple discovery processes can be stored in the same story.

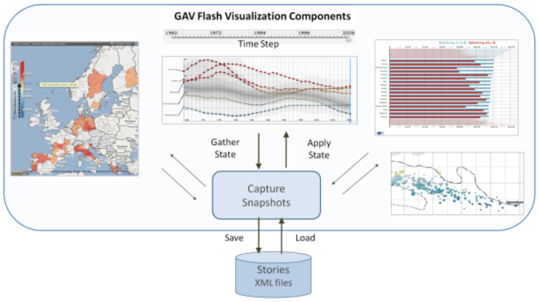


Figure 3. The snapshot tool creates series of captures during an exploration process that form the story.

B. Publish Vislets

Publisher is the tool that takes a story created by World eXplorer and generates the HTML code that can be used to publish a Vislet on the web (Figure 4). A Vislet is a standalone Flash application (widget) that has been assembled from low-level GAV Flash components in a class library (Figure 5).

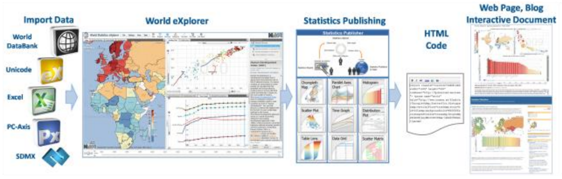


Figure 4. The Exploration and Publishing process

When the user has a finished story that they want to publish they login in to the publisher portal and upload their story to the server. In publisher the user first chooses the layout of the Vislet to be used for the visualizations. This may be everything from a single view layout up to a divided area with multiple visualizations linked to each other. The user then chooses for each view the appropriate visualization to be used. After that the user has some extra options for functionality within the Vislet. When the user is ready the publisher then deploys the story as HTML code, which has all required links to geographic and data sources. The code can then be pasted into the user's favourite web site. A Vislet is created and can now be opened inside the reader's Web browser and communicates the story.

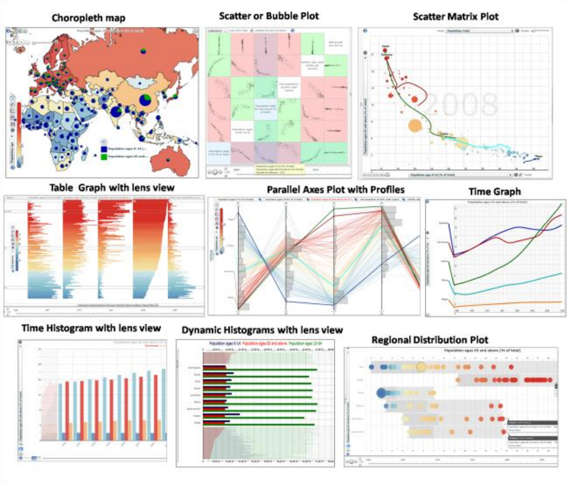
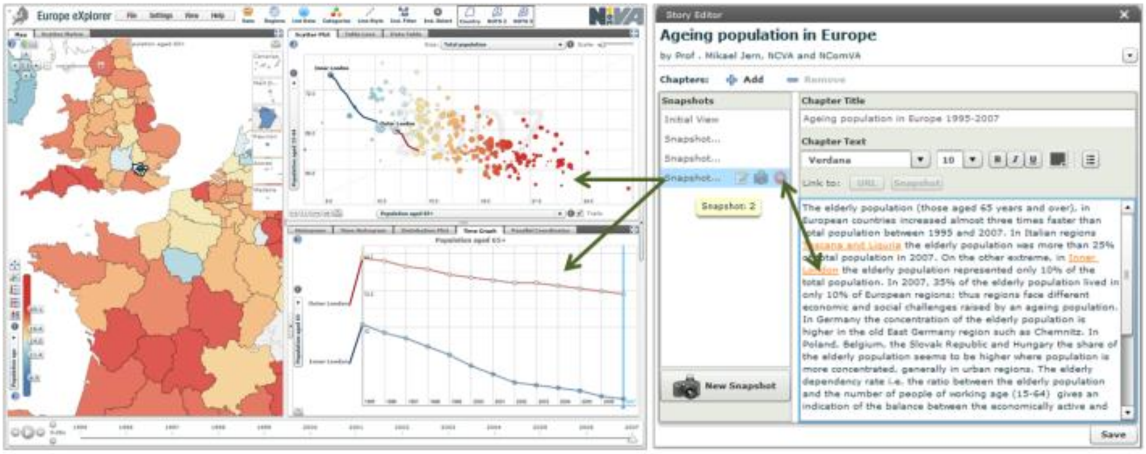

Choropleth map
Scatter or Bubble Plot
Scatter Matrix Plot
2008
Table Graph with lens view
Parallel Axes Plot with Profiles
Time Graph
Time Histogram with lens view
Dynamic Histograms with lens view
Regional Distribution Plot

Figure 5. Example of standalone Vislets for publication

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4



Ageing population in Europe

by Prof . Mikael Jern, NCVA and NComVA

Chapter Title: Ageing population in Europe 1995-2007

The elderly population (those aged 65 years and over), in European countries increased almost three times faster than total population between 1995 and 2007. In Italian regions Toscana and Liguria the elderly population was more than 25% of total population in 2007. On the other extreme, in Inner London the elderly population represented only 10% of the total population. In 2007, 35% of the elderly population lived in only 10% of European regions: thus regions face different economic and social challenges raised by an ageing population. In Germany the concentration of the elderly population is higher in the old East Germany region such as Chemnitz. In Poland, Belgium, the Slovak Republic and Hungary the share of the elderly population seems to be higher where population is more concentrated, generally in urban regions. The elderly dependency rate i.e. the ratio between the elderly population and the number of people of working age (15-64) gives an indication of the balance between the economically active and

Figure 6. Story Editor with the associate Snapshot pointing to "Inner London" in all three views in this regional European visual analytics. The author has access to all stored snapshots in a list on the left side and can easily go to one of them and change the parameters. The Snapshot is linked to the metatext "On the other extreme, in Inner London the elderly population represented only 10% of the total population". Clicking on this Snapshot in a published Vislet immediately initiates all eXplorer views to the state given by the Snapshot.

C. Interactive document

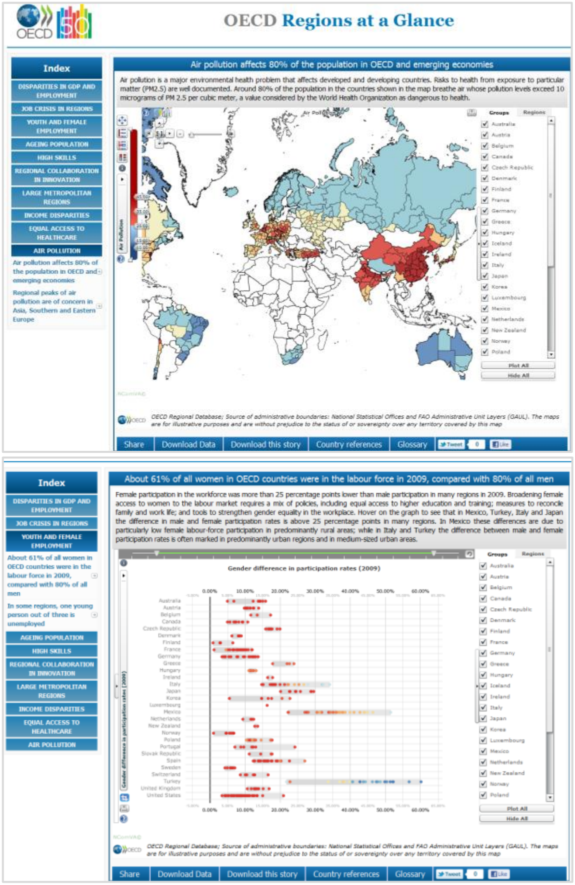
An interactive document is based on a wider storytelling concept, where Vislets play the role of images and figures. This adds another depth to the report or publication by making diagrams interactive, which allows the reader to reach a deeper understanding and further explore the subject. Readers can run animations, change indicators and view more details on specific figures.

One such publication where an interactive document has supplemented the normal paper version is Organisation for Economic Co-operation and Development (OECD) Regions at a Glance (RaG) [18]. This bi-annual report now has an interactive version where selected chapters have been transferred from the publication to an interactive state and published on the web (Figure 7).

The OECD analysts created stories and snapshots from the RaG data, wrote new or edited existing text and collected interesting links. The stories where used to create chapters in the interactive document, where visualizations reside together with the analysts text, relevant links, the corresponding chapter in the publication as a pdf and access to the source data

The interactive document platform used in this case is built to be simple, flexible and portable. Chapters are stored as eXtensible Markup Language (XML) files with lists of their internal sections, the text to display, and the settings for the visualization. These chapters are transformed to html using EXtensible Stylesheet Language (XSLT). This setup might sound simplistic, as using a data base would allow for more advanced and easier updates, but the choice was made with portability in mind. The aim was to allow for very simple installation, no creation of data bases and very limited configuration. In fact, the only setup needed is writing base urls and switching between using Hypertext Preprocessor (PHP) and Active Server Pages (ASP) for running the XSLT.

.



Air pollution affects 80% of the population in OECD and emerging economies

Air pollution is a major environmental health problem that affects developed and developing countries. Risks to health from exposure to particular matter (PM2.5) are well documented. Around 80% of the population in the countries shown in the map breathe air whose pollution levels exceed 10 micrograms of PM 2.5 per cubic meter, a value considered by the World Health Organization as dangerous to health.

Air pollution affects 80% of the population in OECD and emerging economies

Regional peaks of air pollution are of concern in Asia, Southern and Eastern Europe

OECD Regional Database; Source of administrative boundaries: National Statistical Offices and FAO Administrative Unit Layers (GAUL). The maps are for illustrative purposes and are without prejudice to the status of or sovereignty over any territory covered by this map

About 61% of all women in OECD countries were in the labour force in 2009, compared with 80% of all men

Female participation in the workforce was more than 25 percentage points lower than male participation in many regions in 2009. Broadening female access to women to the labour market requires a mix of policies, including equal access to higher education and training; measures to reconcile family and work life; and tools to strengthen gender equality in the workplace. Hover on the graph to see that in Mexico, Turkey, Italy and Japan the difference in male and female participation rates is above 25 percentage points in many regions. In Mexico these differences are due to particularly low female labour-force participation in predominantly rural areas; while in Italy and Turkey the difference between male and female participation rates is often marked in predominantly urban regions and in medium-sized urban areas.

About 61% of all women in OECD countries were in the labour force in 2009, compared with 80% of all men

In some regions, one young person out of three is unemployed

OECD Regional Database; Source of administrative boundaries: National Statistical Offices and FAO Administrative Unit Layers (GAUL). The maps are for illustrative purposes and are without prejudice to the status of or sovereignty over any territory covered by this map

Figure 7. OECD interactive visualization report

Although the choice was made to not use an elaborate system for this publication, different approaches are available. The Vislet technology can be fully integrated into any existing content management system to achieve similar results.

The main benefit of making the report interactive is that it enables the user to visualize the data that is of interest to them, and also to show the variation over time using animation. The analysts at OECD have chosen snapshots that are of interest, but the user can interact with the visualization and change the indicators and select what they find interesting, and thus enhance the user experience.

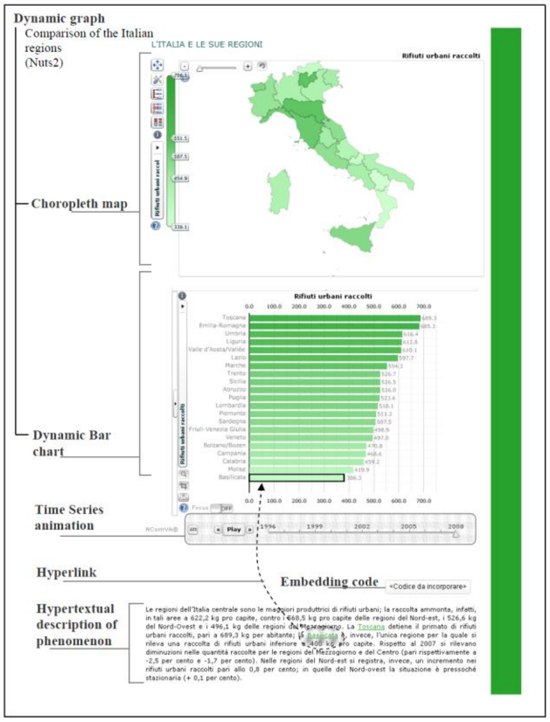
IV. CONCLUSION

The primary objective of the introduction of visualization is to avoid boring statistics. We want official statistics to be exciting so that they invade people's minds and create knowledge, thus allowing users to apply new insights as a basis for decisions. In order to do this the visualization should highlight structures and patterns in the data and allow the users to play around and test their own hypotheses and ideas. It is well known that interactive web based maps speak to the minds of many people. In addition to maps showing the data, we want to highlight interesting correlations through our snapshot mechanism between several indicators across the geographic areas, as well as dynamics over time. If users are to devote interest to statistics, gain knowledge, and use the information for decisions, the statistics must clearly convey a message. The visual analytics storytelling technique applied to statistics visualization therefore is essential.

Focus has historically been on text and picture comprehension but given the explosion of representations made available since the introduction of graphical interfaces, the field now considers all forms of representation including but not limited to, text, pictures, graphs, diagrams, concept maps, animations, equations, virtual reality, information-and scientific visualization, haptics, multimedia, hypermedia, and simulations. Research on learning, when using these aids, is essential. There is research of learning with multimedia environments in different experimental studies but there is hardly any research done of this in real school contexts i.e., in a socio cultural perspective. NCVA is also engaged in a project about social science education in Swedish schools [19].

Within an international perspective our research builds upon collaborating work with many world-leading statistical organizations including the Eurostat, European Commission, OECD [20], [21], Statistics Denmark, Finland, Sweden [22] providing innovative geovisual analytics technology to these organizations. We have also been involved in the development of visualization for the PISA2009 profiles [23]. The national Italian bureau (ISTAT) provides another very interesting and sophisticated learning material [24] about the development and progress in Italian regions in (Figure 8).

To ensure that users would come forward and contribute to a good implementation of geovisual analytics methods applied to statistics, NCVA established a small group consisting of users with different profiles including experts from international and national statistics organizations (OECD, ISTAT, Sweden Statistics), regional researchers, municipal planners and in 2011 was extended to also include school teachers and students. The group met first time already in 2010 and was invited to test and evaluate solutions based on Statistics eXplorer. The feedback has been very positive, as most user categories found that services along these lines would greatly help them understand and use the proposed visualization and storytelling methods. They also gave useful input regarding aspects of the web-enabled presentation, some asking for additional and more complex methods and interactive features, others for simplifications. This confirms the expectation that we must have different levels of service for different segments of users. This type of customization is possible to manage with our fundamental component based tools to assemble the eXplorer applications.



Le regioni dell'Italia centrale sono le maggiori produttrici di rifiuti urbani; la raccolta ammonta, infatti, in tali aree a 622,2 kg pro capite, contro i 668,5 kg pro capite delle regioni del Nord-est, i 526,6 kg del Nord-Ovest e i 496,1 kg delle regioni del Mezzogiorno. La Toscana detiene il primato di rifiuti urbani raccolti, pari a 689,3 kg per abitante; la Basilicata è, invece, l'unica regione per la quale si rileva una raccolta di rifiuti urbani inferiore a 400 kg pro capite. Rispetto al 2007 si rilevano diminuzioni nelle quantità raccolte per le regioni del Mezzogiorno e del Centro (pari rispettivamente a -2,5 per cento e -1,7 per cento). Nelle regioni del Nord-est si registra, invece, un incremento nei rifiuti urbani raccolti pari allo 0,8 per cento; in quelle del Nord-ovest la situazione è pressoché stazionaria (+ 0,1 per cento).

Figure 8. ISTAT – the Italian national statistical bureau has developed this innovative Vislet web site providing an interesting example of a sophisticated learning material.

This test and comprehensive evaluation practice has then continued over several new improved versions. Here are some key features that are highlighted by our partners:

- Mechanism for Storytelling and Snapshots for creating dynamic Web documents;
- Identify what's meaningful in the data in ways that your eyes can see and brain can understand and keep focused on what's important;
- Support large spatio-temporal and multi-dimensional regional data using a data cube;

- Established information visualization technology for multivariate data is adapted and customized to the statistics domain e.g., histogram with fish eye technique, table lens, parallel coordinates with profiles, scatter matrix linked to the scatter plot, choropleth map linked to a treemap;
- Data is simultaneously explored through multiple-linked and coordinated views;
- Map layer architecture - overlay several types of maps such as coloured statistical regions, country boundaries, background maps e.g., Google, pie chart – control transparency level for each layer;
- Dynamic state-of-the-art colour legend with integrated statistical methods such as percentiles;
- Visual inquiry and conditioned statistics filter mechanisms highlighting important discoveries or detecting outliers;
- Interactive time animation simultaneously visualized in all active views;
- Screen space usage is optimized for visualization– no unnecessary visible GUI panels;
- Support for categorical data and visual classification of, for example, urban, rural and intermediate regions;

ACKNOWLEDGMENT

This research and case studies were carried out by NCVA in close collaboration with OECD and Statistics Sweden who supplied data and comprehensive evaluation of the statistical storytelling system. The research is in part supported by funding from the Swedish Agency for Innovation Systems (VINNOVA), the “Visualization Program” coordinated by the Swedish Knowledge Foundation and the Swedish foundation “Marcus och Amalia Wallenbergs Minnesfond” for the educational project VISE "Visual Storytelling in Education". The authors thank the entire research team at NCVA, Linkoping University for valuable contributions and Linnéa Stenliden VISE.

REFERENCES

[1] Andrienko G., Andrienko N., Demsar U., Dransch D., Dykes J., Fabrikant S. I., Jern M., et al. (2010). “Space, Time and Visual Analytics”. International Journal of Geographical Information Science, 24(10), 1577-1600. Taylor & Francis.

[2] Jern M., Thygesen L., Brezzi M. “Storytelling – “How to visualize statistics”, Reviewed paper, NTTS 2011 (New Techniques and Technologies for Statistics), international scientific conference on the impact of new technologies on statistical collection, production and dissemination systems, Brussels, February 2011.

[3] Maceachren A. M., Brewer I., Steiner E. (2001). “Geovisualization to mediatate collaborative work”. Tools To Support Different-Place Knowledge Construction and Decision-Making. Environment, (3), 2533-2539. ICA.

[4] Jern M., “Smart Documents for Web-Enabled Collaboration”, Published in “Digital Content Creation”, Vince J. A. and R. A. Earnshaw (Eds) Springer Verlag, June 2001

[5] Carr D., White D., and MacEachren A.M., “Conditioned choropleth maps and hypothesis generation”. In Annals of the Assoc of Am. Geographers, vol 95, no. 1, 2005.

[6] Robinson A., “Re-Visualization”: Interactive Visualization of the Progress of Visual Analysis, workshop proceedings, VASDS. 2006.

[7] Baumann A., The design and implementation of Weave: A session state driven, web-based visualization framework, University of Massachussets Lowell, 2011

[8] Vie´gas FB., Wattenberg M., Ham FV., Kriss J. and McKeon M. “Many eyes: A site for visualization at Internet scale”. IEEETrans Vis Comput Graph 2007; 13(6): 1121–1128

[9] World Statistics eXplorer with ageing population in the world: http://www.ncomva.se/v4/world/alt.html#story=1

[10] Jern M., Rogstadius J., Åström T., and Ynnerman A: ”Visual Analytics presentation tools applied in HTML Documents”, Reviewed proceedings, IV08, London, July 2008, published by IEEE Computer Society.

[11] OECD web site for visualization of regional development: http://www.oecd.org/GOV/regionaldevelopment

[12] M. Jern, “Collaborative Educational GeoAnalytics applied to large statistics temporal data”, Reviewed proceedings, CSEDU 2010, Conference on computer supported education, Valencia, April 2010

[13] Jern M, “Educating students in official statistics using embedded geovisual analytics storytelling methods”, Reviewed Proceedings in Eurographics 2010, Norrköping, May 2010.

[14] Stenliden L., Jern M., “Educating official statistics using geovisual analytics storytelling methods”, Reviewed proceedings, International Technology, Education and Development Conference INTED, Valencia, 2010.

[15] Kinzel M., Wright D., “Using Geovisualizations in the Curriculum: Do Multimedia Tools Enhance Geography Education?” Paper Number 1290, Environmental Systems Research Institute Education User’s Conference, 2008.

[16] Quan H., Lundblad P., Åström T., and Jern M., “A web-enabled visualization toolkit for geovisual analytics”. Journal Information Visualization. Published online before print November 9, 2011, doi: 10.1177/1473871611425870.

[17] Quan H., Lundblad P., Åström T., and Jern M., “A Web-Enabled Visualization Toolkit for Geovisual Analytics Visualization and Data Analysis”, Awarded best paper, SPIE: Electronic Imaging Science and Technology, Visualization and Data Analysis, Proceedings of SPIE, San Francisco Jan 2011

[18] OECD interactive report of regional development http://rag.oecd.org/

[19] Stenliden L., Jern M., “Visual Storytelling – Understanding and Knowledge in Education”, Awarded best paper, International Symposium on Academic Informing Science & Engineering: isAISE 2011 in the context of The 2nd International Multi-conference on Complexity, Informatics and Cybernetics: IMCIC 2011, Orlando, Florida, USA. ISBN-13: 978-1-936338-21-4.

[20] Jern, M., Thygesen L., Brezzi M., “A web-enabled Geovisual Analytics tool applied to OECD Regional Data”, Reviewed Proceedings in Eurographics 2009, Munich, March 2009

[21] OECD web site for visualization of regional statistics: http://stats.oecd.org/OECDregionalstatistics/

[22] The Statistic Atlas from Statistics Sweden http://www.scb.se/Kartor/Statistikatlas_KN/index.html

[23] OECD PISA 2009 profiles by country/economy: http://stats.oecd.org/PISA2009Profiles/

[24] Our italy, by the national Italian bureau ISTAT: http://noi-italia.istat.it/

Characterization of the Wikipedia Traffic

Antonio J. Reinoso
Libresoft Research Group (URJC)
Fuenlabrada (Spain)
ajreinoso@libresoft.es

Rocío Muñoz-Mansilla
Department of Automation and Computer Science (UNED)
Madrid (Spain)
rmunoz@dia.uned.es

Israel Herraiz
Department of Applied Mathematics and Computing (UPM)
Madrid (Spain)
israel.herraiz@upm.es

Felipe Ortega
Libresoft Research Group (URJC)
Fuenlabrada (Spain)
jfelipe@libresoft.es

***Abstract*—Since its inception, Wikipedia has grown to a solid and stable project and turned into a mass collaboration tool that allows the sharing and distribution of knowledge. The wiki approach that basis this initiative promotes the participation and collaboration of users. In addition to visits for browsing its contents, Wikipedia also receives the contributions of users to improve them. In the past, researchers paid attention to different aspects concerning authoring and quality of contents. However, little effort has been made to study the nature of the visits that Wikipedia receives. We conduct such an study using a sample of users' requests provided by the Wikimedia Foundation in the form of Squid log lines. Our sample contains more that 14,000 million requests from users all around the world and directed to all the projects maintained by the Wikimedia Foundation, including different editions of Wikipedia. This papers describes the work made to characterize the traffic directed to Wikipedia and consisting of the requests sent by its users. Our main aim is to obtain a detailed description of its composition in terms of the percentages corresponding to the different types of requests making part of it. The benefits from our work may range from the prediction of traffic peaks to the determination of the kind of resources most often requested, which can be useful for scalability considerations.**

***Keywords*-Wikipedia; Traffic characterization.**

I. Introduction

Wikipedia stands as the most successful wiki-based project and provides a vast compilation of contents related to all the knowledge areas. Furthermore, the Wikipedia underlying philosophy promotes the collaboration and participation of users in the production of pieces of knowledge that will remain available for the whole community. This new paradigm for knowledge generation has attracted great attention and has propitiated the consolidation of Wikipedia as a mass collaboration tool. Such acceptation can be regarded just from the continuously increasing number of visits to the different Wikipedia domains that places its web site within the six most visited ones all over the Internet [1].

Regarding its contents, Wikipedia is divided in approximately 270 [2] editions that correspond to the same number of languages. All these editions sum up to 19 million articles, which correspond to encyclopedic entries about particular subjects, dates or people. Wikipedia articles address topics corresponding to traditionally academic disciplines as well as to cultural, sportive or artistic manifestations. In addition, they also deal with highly topical subjects, biographies from live persons or topics related to general public entertainment.

In respect to the audience, Wikipedia editions receive approximately 13,500 million visits a month. This observation can be considered as a good indicator of its acceptance and popularity among users. Such number of visits constitute an absolute challenge in terms of management of requests and content delivery. Concerning this topic, several re-arrangements and re-organizations had to be made in the supporting architecture to meet the scalability demands derived from its rise in popularity and users' participation.

As a result of this relevance, Wikipedia has evolved into a subject of increasing interest for researchers [3]. In this way, different quantitative examinations about its articles, authors, visits or contributions have been undertaken [4], [5]. However, most of the previous research involving Wikipedia deals with the quality and reliability of its contents ([6], [7], [8]) or study its growing tendency and evolution [9], [10]. By contrast, very few studies [11], [12], [13] have been devoted to analyze the manner in which users interact and make use of Wikipedia.

Therefore, this paper aims to analyze the different kind of requests submitted to Wikipedia by its users in a effort to determine both quantitative and qualitative features of such traffic. The major benefits from our study may range from a detailed characterization of the requests sent to the Wikipedia supporting architecture to the forecasting of systems' overload during stress peaks. In addition, proper knowledge about most requested resources may lead to systems improvements concerning the delivery management policy. Finally, we also provide several comparisons amongst the different Wikipedia editions in order to assess differences or trends relative to particular editions. Moreover, we also outline those evolutions that do not fit the general tendency resulting from the observation of all the received requests.

Our analysis focuses on the largest Wikipedia editions in terms of their number of both articles and requests. In addition, we have analyzed the traffic during a whole year (2009) to avoid temporarily localized events. Our main data source consists in users' requests that are stored by special Squid servers that are deployed by the Wikimedia Foundation to deal with all the incoming traffic to its several projects. In this way, information about each individual request is registered in the form of a log line whose fields are later processed by an ad-hoc Java application. This application filters the requests targeting to Wikipedia contents or services and classifies them for countability purposes.

The rest of the paper is structured as follows: Section II presents the data sources used for this study. Section III explains the filtering process for the data sample and the information that can be extracted out of it. After this, Section IV presents the results, and finally Section V concludes this paper and proposes some ideas for further work.

II. The data source

This section aims to describe the information sources used in our study and constituting the main data feeding to perform our analysis. Visits to Wikipedia, in a similar way to any other Internet site, are issued from users' browsers in the form of URLs. These petitions are registered by the Wikimedia Foundation Squid servers in the form of log lines once the requested contents have been served.

Squid servers are a special kind of servers performing web caching that are used by the Wikimedia Foundation as the first layer in its Content Delivery Network. They manage all the traffic directed to Wikipedia as well as to the rest of wiki-based projects. Squids register every responded petition as log lines and a sample of them is sent to universities and research centers.

Squids commonly work as proxy servers performing web caching. In this way, they cache contents previously browsed to make them locally available in the case that requests for the same contents are issued. This results in a significant decrease of the bandwidth consumption and in a more efficient use of the underlying network. Furthermore, Squid servers may also be used to improve web servers by caching the contents repeatedly requested to them. Squid servers are said to work as reverse proxy servers because they try to answer the incoming requests with the cached contents. When successful, this approach avoid the participation of any other system in the delivery of the requested contents. Particularly, this prevents the operation of database or web servers purportedly placed behind them.

In the case of the Wikimedia Foundation, two layers of Squid servers are placed in front of its Apache and database servers. In this way, most of the requested content is directly served from the Squid subsystem without involving any of the other servers. As the Wikimedia Foundation maintains several wiki-based projects, such as Wikipedia, Wikiversity or Wikiquote, the Squid layers have to deal with all the incoming traffic directed to these projects.

Currently, there are two large Squid server clusters: a primary cluster (located in Tampa, Florida) and another secondary cluster (located in Amsterdam) that only performs web caching. These Squids servers usually run at a hit-rate of approximately 85% for text and 98% for media using CARP (Cache Array Routing Protocol) [14]. Users' requests are firstly routed to one of the Squid clusters using a DNS balancing policy.

However, all the contents requested by users are not cacheable. The pages sent to registered and logged-in users, for example, cannot be cached as they include customized parts as the users' nicknames or, even, personalized options for page displaying such as skins or templates.

Squid systems log information about each served request disregarding whether the answer could have been found in the cache or, on the contrary, it was a tailored page built up by web servers. Every Squid server packages and sends its log lines to a central aggregator host. Here, there is a program in charge of their reception that, in addition, sends them to the set of registered log processors. Basically, a log processor consists either in a file processor, that writes lines to a file, or in a pipe processor, that sends them to a specific command trough a pipe. Both of them use a sampling factor to determine the next line to be written or piped. In turns, another program does the opposite operation and picks the lines to send them through a UDP packet stream. This is how Wikimedia Foundation Squid log lines finally reach our storage systems.

Each log line from a Wikimedia Squid server corresponds to a served user request and constitutes a really valuable data feed because, among several other information, it includes the URLs submitted by the user along with the date at witch the corresponding content was sent in response.

III. Methodology

The analysis presented here is based on a sample of the traffic directed to all the Wikimedia Foundation wiki-based projects during 2009. The sampling factor used for generating our data feed was 1%, which means that we received one in every hundred requests composing the traffic to the several projects maintained by the Wikimedia Foundation. In general terms, more than 14,000 million log lines have been parsed and filtered for this study.

To begin with, we had to separate the requests directed to Wikipedia from the ones targeting to projects like Wikiquote, Wikiversity, etc. In addition, we have only considered consolidated and assiduous Wikipedias in order to focus on highly active editions. Specifically, we have analyzed the requests corresponding to the ten top-most editions regarding their number of, both, articles and visits. These editions are the German, English, Spanish, French, Italian, Japanese, Dutch, Polish, Portuguese and Russian ones.

The streaming made up of the log lines from the Wikimedia Foundation Squid systems is daily rotated in such a way that lines corresponding to different days are separated in different files. Once stored, log lines are completely available for their processing using an ad-hoc java written application: the *WikiSquilter* tool [15]. The processing consists in a parsing phase devoted to extract the relevant information fields from the log lines. Then, these elements are filtered in order to determine what lines correspond to requests considering of interest according to the directives of the driven analysis. Finally, data related to filtered requests are normalized and stored in a relational database for further examinations.

Log lines received from the Wikimedia Foundation offer a valuable information by themselves though they do not contain specific fields with the necessary data to conduct our analysis. However, these data can be obtained from the URLs submitted by users when they send a request. In this way, URLs have to be parsed to look for the precise elements involved in the characterization process. In particular, there are elements that can be easily extracted from requests such as the following ones:

1) The Wikimedia Foundation project, such us Wikipedia, Wiktionary or Wikiquote, to which the URL is directed.
2) The corresponding language edition of the project.

For the rest of information elements, the parsing process relies on the use of regular expressions to determine the syntactical structure of requests and, consequently, their purported type. In particular, we aim to characterize users' petitions consisting in:

1) Visits, intended as requests for browsing (reading) Wikipedia articles that do not convey any other action.
2) Any action such as previews, edit historical reviews but excluding edits and searches that are treated separately.
3) edits, sent to modify Wikipedia contents that cause the issue of write operations to the database.
4) searches, looking for articles related to a certain topic.
5) api calls, that request any of the built-in functionality offered by the mediawiki software.
6) skin/css requests, that demand customized elements or choices used in the visualization and presentation of Wikipedia contents.
7) media wiki extensions, that are requests for extensions added to provide new functionalities through third-party code ready to be set up together with the mediawiki core.

The filter process consists in assessing whether each analyzed URL is considered significant for our analysis. This is done by checking whether the information elements it contains correspond to the ones in which our work is focusing on. The filter implementation is realized on the basis of a hash structure holding the information elements considered of interest for the analysis as well as their corresponding normalized database codes to be used in the insert operation issued to database management system.

In general terms the application has been designed and developed with strong adherence to the principles of efficiency, robustness and accuracy. However, flexibility and extensibility directives have been also reinforced. Efficiency is gained through several elements such as multithreaded design and filter's O(1) complexity derived of the hash basis. Application's robustness has been achieved by means of the capability of detecting malformed URLs. Flexibility makes the application suitable of being used with whatever log lines with the only requirement of specifying in the corresponding XML file the elements to be parsed and filtered. The software architecture of the application allows to easily include new services that can even involve new data to be processed, so extensibility has been also considered.

IV. Analysis and results

This section provides a quantitative analysis of the traffic composition in the aim of providing an adequate characterization of the requests directed to Wikipedia. This kind of analysis may contribute to describe the way in which Wikipedia is being utilized by its community of users. In addition, our results may serve as an estimation of the operational overload for the systems in charge of supporting the different Wikimedia Foundation wiki-based projects and, particularly, Wikipedia.

Therefore, we present here the characterization of the different types of requests composing the traffic to the Wikipedia editions under study. Furthermore, we are also presenting information related to the general traffic to all the Wikimedia Foundation projects. Traffic information is always computed in terms of number of requests, disregarding, by the moment, considerations about amount of information or transference rates. In addition, we usually present the daily averaged number of requests when larger time units, such as months, are considered in order to avoid the introduction of biased perceptions due to the differences in the number of days. We have to note that technical problems have prevented us from obtaining the traffic information of all the days from 2009. However, we have only failed to receive the traffic of just 4 days, which is an absolute success in terms of the reliability of our receiving infrastructure.

First of all, we consider of interest to determine how the overall traffic to the Wikimedia Foundation is distributed among its different projects during 2009. This is shown in Table I, which provides the percentages of the total traffic corresponding to each particular project. As it is clearly seen, the largest percentage corresponds to the requests

WMF project	Percentage of traffic attracted
Wikipedia	49.47%
Wikiversity	0.03%
Wikibook	0.23%
Wiktionary	0.52%
Wikiquote	0.16%
Species	0.01%
Wikinews	0.06%
Wikisource	0.13%
Commons (images)	1.26%
Uploaded resources	46.72%
Other	1.41%

Table I
TRAFFIC DIRECTED TO EACH WIKIMEDIA FOUNDATION PROJECT AND TO PREVIOUSLY UPLOADED RESOURCES.

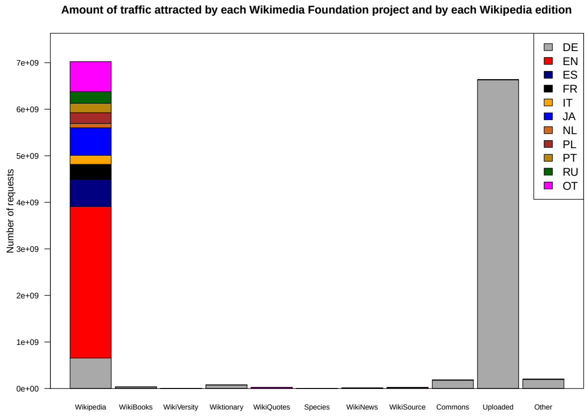


Figure 1. Amount of traffic corresponding to each Wikimedia Foundation project and to each edition of Wikipedia during 2009.

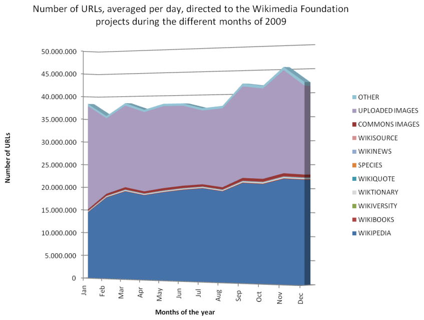


Figure 2. Evolution of the daily averaged number of requests to each Wikimedia Foundation project in every month of 2009.

for Wikipedia articles (49.7%). Interestingly, almost the remaining half of requests (46.72%) corresponds to images and other multimedia resources uploaded to the platform to be referenced not only from Wikipedia articles but also from articles belonging to the rest of wiki-based projects. All together, these two types of requests add up to the 96% of all the traffic received by the Wikimedia Foundation servers. Figure 1 shows the relevance of both kind of requests in the traffic and also includes the amount of it corresponding to each Wikipedia edition. As it is shown, the English Wikipedia (EN, in red) attracts much more traffic than any other edition followed by the German (DE, gray), the Spanish (ES, dark blue) and the Japanese (JA, blue) ones.

Figure 2 presents the monthly evolution of the traffic directed to all the Wikimedia Foundation projects during 2009. The vertical axis shows the daily average of requests corresponding to each particular project and to common resources, mainly images, requested by users. In order to adequately examine these figures, it is important to remark that they correspond to the daily average of the sample we receive, which is the 1% of the total traffic, so real ones would be, for example, 30 * 100 times higher in the case of months having 30 days. From Figure 2 we can also compare the monthly evolution of the traffic to Wikipedia in respect to the total traffic to all the supported Wikimedia Foundation projects. As it is shown, though both traffics seem to follow quite similar monthly evolutions, there are some differences such as the tendencies observed in February that indicate that the total traffic decreased whereas the number of requests to the set of Wikipedias increased.

From here on, we aim to characterize only the traffic corresponding to the Wikipedia project. In this way, our first objective is to determine the number of requests directed to each one of its editions and, particularly, to the ones considered in this work. Thus, Figure 3 shows the distribution of the requests to Wikipedia over its different editions in every month of 2009. The English Wikipedia is still the most popular, and receives a volume of traffic much higher than the other editions. Besides this, we considered of interest to aggregate the daily average of the traffic to each Wikipedia edition throughout the entire 2009 and to present their corresponding percentages in respect to the total traffic to the Wikipedia project. Table II presents this information. As we can see, the considered editions attract more than 91% of the total traffic to Wikipedia. This is important in terms of the relevance of the considered set of editions. The evolution of the daily average of requests for each particular edition in the different months of 2009 is presented in Figure 4. As it is shown, not all the Wikipedias follow the same distribution of traffic over time, which can mean different temporal patterns of use.

We can also compare the evolution of the traffic to the different editions of Wikipedia with the progression of their respective sizes. Larger Wikipedias may attract a higher number of requests as a result of their purportedly bigger supporting community. However, this is not always true according to the Figures 5 and 6 which present, respectively, the amount of traffic attracted by each Wikipedia in every month of 2009 and their sizes expressed in number of articles during the same months.

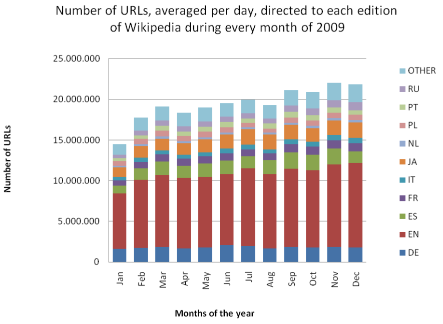


Figure 3. Comparison of the traffic directed to each edition of Wikipedia in every month of 2009.

Wikipedia edition	Daily average of attracted traffic	Percentage
DE	21,767,176.73	9.40%
EN	108,407,534.61	46.45%
ES	19,336,747.61	8.25%
FR	10,622,527.01	4.54%
IT	6,516,987.21	2.79%
JA	19,591,570.27	8.38%
NL	3,128,496.65	1.34%
PL	7,628,743.39	3.30%
PT	6,755,424.08	2.87%
RU	8,269,484.01	3.51%
REST	21,467,547.49	9.17%

Table II
AGGREGATED DAILY AVERAGED NUMBER OF REQUESTS ATTRACTED BY EACH CONSIDERED EDITION OF WIKIPEDIA DURING THE WHOLE 2009. THE TRAFFIC CORRESPONDING TO THE REST OF DISREGARDED EDITIONS IS PRESENTED SUMMARIZED UNDER THE 'REST' ENTRY.

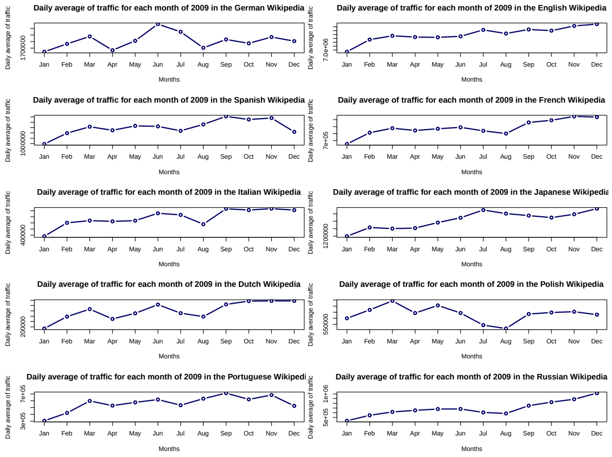


Figure 4. Evolution of the daily averaged traffic directed to each edition of Wikipedia over the different months of 2009.

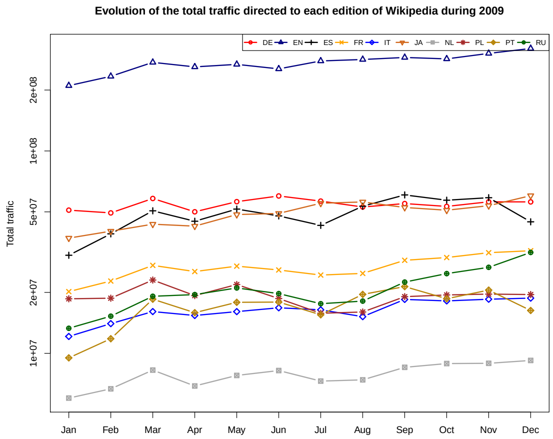


Figure 5. Monthly evolution of the total traffic directed to each edition of Wikipedia throughout 2009.

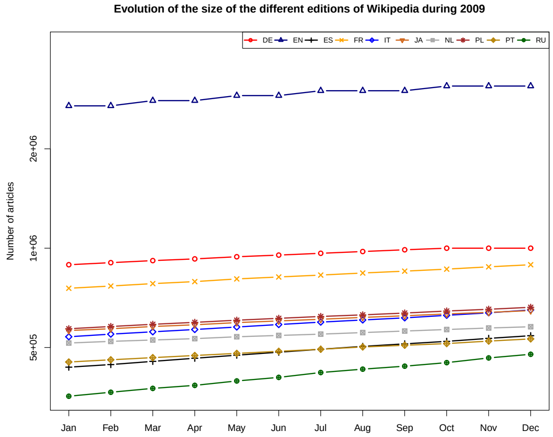


Figure 6. Monthly evolution of the size of the different editions of Wikipedia throughout 2009.

Considering that the English and the Russian Wikipedias are, respectively, the largest and the smallest ones, the same is not valid for the amount of traffic . The case of the Spanish Wikipedia is even more curious because in spite of being situated among the three editions with lesser volumes of articles, regarding its traffic, it ranges from the fourth to, even, the second most requested edition. This finding is specially interesting because it proves that resources related to storage and traffic management are not proportional at all, what should be considered particularly in scalability issues.

Surely, it is more interesting to obtain a characterization of the traffic directed to each edition of Wikipedia. This information could be interpreted as an approximation to the use given to each Wikipedia edition by its corresponding community of users. So,Table III shows the percentage of traffic directed to each Wikipedia edition that consist in visits to articles, requests for edit operations, actions such as history reviews or pre-visualizations performed on articles, search operations, css files used to present tailored pages or, even, the Wikipedia icon itself.

From Table III it is interesting to note the high percentage of requests corresponding exclusively to visits as well as to elements related to the presentation and visualization of the

Ed.	Visits to articles	Actions (exc. edit & search op.)	Edit op.	Search op.	Api calls	Skins /css	icons	mw ext.	Undet.
EN	21.51%	22.52%	0.27%	4.75%	6.53%	34.62%	4.38%	3.47%	6.95%
DE	16.54%	20.87%	0.23%	4.09%	7.69%	30.74%	3.46%	14.72%	5.98%
ES	13.58%	33.90%	0.31%	4.12%	6.02%	32.13%	3.68%	3.89%	6.80%
FR	18.24%	23.15%	0.33%	4.00%	6.05%	36.87%	4.42%	4.23%	7.04%
IT	19.80%	21.81%	0.43%	4.44%	5.77%	37.57%	4.49%	3.07%	9.69%
JA	20.69%	25.15%	0.37%	4.22%	3.95%	36.01%	4.19%	2.81%	9.22%

Table III
CHARACTERIZATION OF THE TRAFFIC DIRECTED TO SOME PARTICULAR EDITIONS OF WIKIPEDIA IN TERMS OF THE PERCENTAGES CORRESPONDING TO DIFFERENT TYPES OF REQUESTS.

requested information. It is also noticeable the extremely low percentage of edits (requests to commit any changes over the contents) that is two order of magnitude less than visits.

Regarding the different types of actions, it is shown that requests consisting in calls to the MediaWiki API (Application Programming Interface), search operations and mediaWiki extensions (pieces of code to add particular functionalities to the wiki engine) present relevant percentages. Again, this information may be useful to set and configure the range of resources dealing with these types of requests. Particular interesting observations such as the low percentage of visits in the German Wikipedia together with the impressively high ratio of requests demanding mediaWiki extensions in this edition deserve deeper research. In the same way, the lower percentage of visits corresponding to the Spanish edition and its higher number of requested actions also deserve thorough efforts.

V. CONCLUSION

In this paper we have shown how the Wikipedia traffic can be characterized to obtain its detailed composition. Furthermore, the analysis of the traffic directed to all the projects maintained by the Wikimedia Foundation indicated that it was composed mainly by requests to Wikipedia, on the one hand, and requests for previously uploaded resources, on the other hand.

When comparing the monthly evolutions corresponding to the traffic directed to the whole set of the Wikimedia Foundation's projects and to the one consisting in the, requests, just, to the contents from Wikipedia, it was found that both evolutions are considerably similar thought they present some differences. In particular, the traffic to Wikipedia presents a temporal distribution with less drops and with a slope slightly more tending to increase. This can be interpreted as a non-stopping raise in the attention attracted by the Wikipedia project. In addition, situations when the number of requests to Wikipedia increases though the general traffic falls might be explained, for example, because of a raise in the demands of articles with less images or graphical contents.

Focusing on the requests to Wikipedia, we have determined how the traffic is distributed among its different editions and how the number of received requests is not related to the editions' sizes. This is particularly interesting as it shows that resources arranged for storage and delivery do not scale with the same ratio. Wikimedia Foundation systems staff may take this fact into consideration when planning the allocation of the different kind of resources to be involved in the management and serving of the requests directed to particular language editions.

In respect to the distribution of the requests over the different months, it is found that, as expected, the traffic generally decreases during the summer months surely associated with holiday periods. In the rest of the months the tendency of the traffic does not fluctuate very much and usually tends to increase.

Finally, the percentages corresponding to the different types of considered requests found in the traffic to each edition have been presented. These results show a high number of visits and solicited actions, both near 20%. This is particularly noticeable because visits may be replied using cached contents provided they were issued by non-logged users. However, actions can never been answered in that way so that they need the participation of database servers and specific software systems depending on the nature of the requested actions.

Regarding some of the differences observed in the percentages of the different kinds of actions found in the analyzed editions, it is clear that further research is needed to find out concrete situations. Particularly, the outstanding amount of traffic concerning visualization options deserves a closer examination as it represents, in average, a third the total traffic to each editions. Depending on whether it corresponds to the established displaying options or not, it impact on the overall performance can be really remarkable.

REFERENCES

[1] "Information about wikipedia.org in alexa," accesed March 2012. [Online]. Available: http://www.alexa.com/siteinfo/wikipedia.org

[2] "Wikimedia statistics," accesed March 2012. [Online]. Available: http://stats.wikimedia.org/EN/Sitemap.htm

[3] “Academic studies of wikipedia,” accesed March 2012. [Online]. Available: http://en.wikipedia.org/wiki/Wikipedia:Academic_studies_of_Wikipedia

[4] J. Voss, “Measuring wikipedia,” in *International Conference of the International Society for Scientometrics and Informetrics : 10th*. ISSI, July 2005.

[5] F. Ortega, J. M. Gonzalez-Barahona, and G. Robles, “The top ten wikipedias: A quantitative analysis using wikixray,” in *Proceedings of the 2nd International Conference on Software and Data Technologies (ICSOFT 2007)*, INSTICC. Springer-Verlag, July 2007. [Online]. Available: http://libresoft.es/downloads/C4_159_Ortega.pdf

[6] Korfiatis, Nikolaos, Poulos, Marios, Bokos, and George, “Evaluating authoritative sources using social networks: an insight from wikipedia,” *Online Information Review*, vol. 30, no. 3, pp. 252–262, May 2006. [Online]. Available: http://dx.doi.org/10.1108\%2F14684520610675780

[7] J. Giles, “Internet encyclopaedias go head to head,” *Nature*, vol. 438, no. 7070, pp. 900–901, December 2005. [Online]. Available: http://dx.doi.org/10.1038\%2F438900a

[8] T. Chesney, “An empirical examination of wikipedia’s credibility,” *First Monday*, vol. 11, no. 11, November 2006. [Online]. Available: http://firstmonday.org/issues/issue11_11/chesney/index.html

[9] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, “Preferential attachment in the growth of social networks: the case of wikipedia,” Feb 2006. [Online]. Available: http://arxiv.org/abs/physics/0602026

[10] D. Spinellis and P. Louridas, “The collaborative organization of knowledge,” *Commun. ACM*, vol. 51, no. 8, pp. 68–73, August 2008. [Online]. Available: http://dx.doi.org/10.1145/1378704.1378720

[11] G. Urdaneta, G. Pierre, and M. van Steen, “A decentralized wiki enginge for collaborative wikipedia hosting,” in *Proceedings of the 3rd International Conference on Web Information Systems and Technologies*, March 2007, pp. 156–163. [Online]. Available: http://www.globule.org/publi/DWECWH_draft2006.pdf

[12] A. J. Reinoso, “Temporal and behavioral patterns in the use of wikipedia,” Ph.D. dissertation, Universidad Rey Juan Carlos, 2011, http://gsyc.es/ ajreinoso/phdthesis.

[13] A. J. Reinoso, J. M. Gonzalez Barahona, F. Ortega, and G. Robles, “Quantitative analysis and characterization of Wikipedia requests,” in *Proceedings of the 4th International Symposium on Wikis*, ser. WikiSym ’08. New York, NY, USA: ACM, 2008. [Online]. Available: http://dx.doi.org/10.1145/1822258.1822302

[14] M. Bergsma, “Wikimedia architecture,” Tech. Rep., April 2007, accesed March 2012. [Online]. Available: http://www.nedworks.org/~mark/presentations/san/Wikimedia\%20architecture.pdf

[15] “The wikisquilter project,” accesed March 2012. [Online]. Available: http://www.alexa.com/siteinfo/wikipedia.org

Generic Wireless Control System – Case study

Petra Seflova
Faculty of Mechatronics, Informatics and Interdisciplinary Studies
Technical university of Liberec, TUL
Liberec, Czech Republic
seflova@microrisc.com

Jiri Pos, Vladimir Sulc
MICRORISC s.r.o.
Jicin, Czech Republic
pos@microrisc.com, sulc@microrisc.com

***Abstract*— Due to the increasing number of various types of wireless networks and their applications there is also an increasing demand for systems enabling to process and display data to be understandable not only to machines but also to people. This article describes a generic wireless control system for data and event logging with Ethernet connectivity and internal web server. It is based on the IQRF platform and an Ethernet gateway by MICRORISC s.r.o. The system is fully user-adaptable by programming wireless module with operating system in C language and a web application using PHP scripts.**

Keywords- IQRF, GW-ETH-01, mysql, sql, php, web, datalogger

I. INTRODUCTION

Current industrial equipment utilizing wireless communication often requires not only data transfer, processing and administration, but also to be displayed in a user-friendly format. Another need may be data accessibility from everywhere.

Example: Wirelessly transferred data between two industrial devices in hexadecimal format is 001F2A3B. This code is understandable for a machine but for a human the following record is more illustrative: Upper limit: 50°, lower limit: 10°.

The task is to design equipment allowing turning of pump vanes controlled either at the pump or from a remote station, and also via Internet. At the same time there is a request to archive all events regarding given equipment. The archiving must proceed automatically within a period specified by the user. The system should be as robust as possible, with low power consumption. The device is placed inside the vane, which is why cables cannot be used for interconnection.

Thus, a wireless solution has been chosen. At present, several wireless platforms are available: ZigBee [8], MiWi [9], IQRF [2], etc.

ZigBee is a wireless communication technology based on the IEEE 802.15.4. standard. It is relatively new, effective from November 2004. It is intended for low power devices in PAN (Personal Area Networks) and low range (up to 75 m). Thanks to multihop ad-hoc routing communication is enabled also for higher distances without direct visibility. ZigBee is primarily used in industry and sensor networks. It works in license-free bands 868 MHz, 902–928 MHz and 2.4 GHz. Transmission rates are 20, 40 and 250 kbit/s.

IQRF is a platform for ***low speed***, ***low power***, *reliable and easy to use wireless connectivity*.

- RF bands 868 MHz and 916 MHz (ISM)
- Based on transceiver modules with built-in operating system (OS)
- Fully open functionality depends just on user-specific application written in C language
- Packet-oriented communication, max. 128 B per packet
- Range up to several hundred of meters per hop, up to 240 hops per packet
- Extra low power consumption: 900 nA or 2 µA standby, 35 µA receiving
- Low bit rate: 1.2 kb/s – 115 kb/s
- No license fee

Based on comparison of available platforms, their features, availability and costs, the IQRF platform has been selected.

This article is separated into the following sections: Section II briefly describing the IQRF platform, Section III introducing a web application to manage data, and Section IV describing a real application SIGMA.

II. IQRF

IQRF is an abbreviation for *intelligent* connectivity using *radio frequency*. It is a complex communication platform with modular components for easy user applications. It was introduced in 2004 [10]. It is intended for reliable packet-oriented low power data transfers, either peer-to-peer or in complex networks. Application domains are telemetry, e.g., AMR, Smart metering, WSN, Smart grids, automation of buildings, e.g., Smart house, HAN and cities (Smart city, street lighting), industry and services. This platform is

ISBN: 978-1-61208-200-4

suitable for almost any electronic equipment where wireless communication is needed, e.g., remote control, monitoring, alarm, displaying of remotely acquired data or connection of more devices to a wireless network. The platform is described in [2].

A. Tranceiver module

IQRF transceiver module (TR) is a basic communication component of the platform, used not only in common end devices but also in all IQRF gateways, routers, etc.

It is an intelligent electronic board with complete circuitry needed for implementation of wireless RF connectivity with several peripherals and interfaces. It includes a microcontroller with a built-in operating system (OS) supporting MESH networking, serial EEPROM memory, voltage regulator (LDO) and optionally also temperature sensor and built-in PCB antenna. Besides general I/Os, SPI running in OS background, I^{2}C and UART standard serials can directly be used. The tiny compact module with highly integrated design, typically in SIM card format 25 x 14.9 mm, requires no external components. Several TR types are available differing in performance, MCU type, memory size, peripherals, antenna options and dimensions.

Due to extra low power consumption, e.g., 35 μA while receiving TRs are suitable even for devices with extreme requirements for battery life. TRs work within the license-free ISM band, 868 MHz in Europe and other countries or 916 MHz in America and other countries. All communication parameters are software selectable.

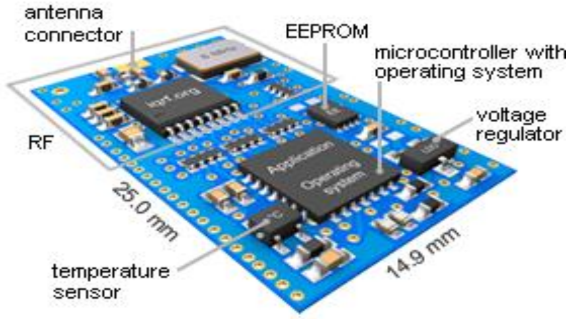


Figure 1. Transceiver module

B. IQRF GATEWAY GW-ETH-01

The GW-ETH-01 Ethernet gateway [1] enables connectivity between an IQRF wireless network and a local LAN and Internet. The gateway has an internal web server allowing communication via HTTPS. Up to two concurrent links are allowed at the same time. Additionally, communication via the UDP transport layer is possible using the IQRF proprietary UDP protocol.

The main GW-ETH-01 parts are 16b MCU, Ethernet driver, EEPROM memory, temperature sensor, IQRF TR module with antenna and backup accumulator.

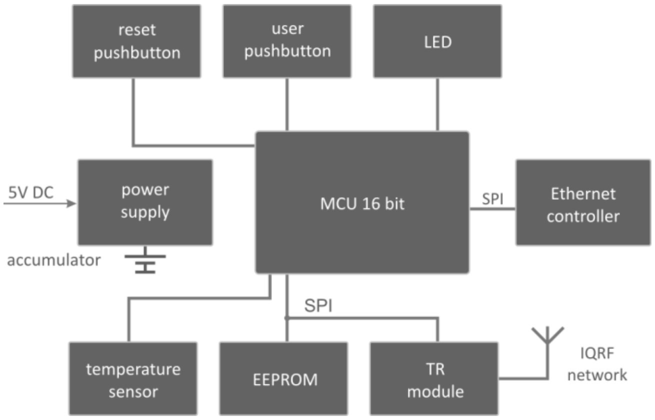
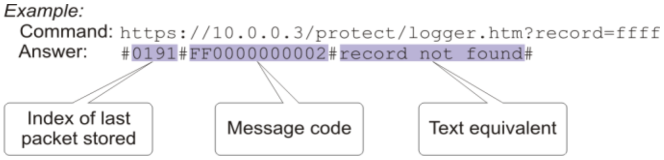


Figure 2. Block diagram of GW-ETH-01

One of communication modes with internal IQRF module is a Datalogger [3]. In this mode, operation depends solely on the application in an internal TR module. All data sent via SPI from the TR module to the GW are stored in a circular buffer in the GW without modifications. Every packet is equipped with numeric code and time stamp. Data can be read from this buffer via the HTTPS interface, or user data can be sent via HTTPS to the TR module.

Buffer parameters:

- Buffer size 7 kB (7 168 B)
- Packet size
  - Serviceable data 1 to 41 (fixed but user selectable)
  - Overhead 8 B



Example:

```
Command: https://10.0.0.3/protect/logger.htm?record=ffff
Answer:  #0191#FF0000000002#record not found#
```

Figure 3. Example command for communication with the Datalogger

GW-ETH-01 can be used, e.g., for:

- Remote monitoring and control
- Data acquisition
- Data storage – datalogger
- Connection of more IQRF networks to single PC
- Access to an IQRF network from more points
- Time synchronization from time servers
- Interface to home automation, etc.

III. WEB APPLICATION

Preconditions:

- TR in GW and TR(s) in user system are programmed according to given needs.
- The GW must be set in the Datalogger mode
- In case of Internet connection, the GW must be registered and connected to IQRF DNS server (Domain Name Server) www.iqrfdns.org [5]. Then the GW polls the DNS server for its own IP (Internet Protocol) address.
- Database server, e.g., MySQL or MS SQL running and a database created for user application on the computer where the application is hosted.
- Web server, e.g., Apache or IIS (Internet Information Server) running on the computer where the application is hosted.

Application is designed as a web application based on *three-layer architecture: Database, Application and Presentation layers*. For security reasons, communication runs on HTTPS protocol *(Hypertext Transfer Protocol Secure*) between GW-ETH-01 and the web application.

The foundation of the application is a *Database layer* using the DBMS (Database Management System) [7]. The *Application layer* is created above the DBMS. It includes most of the application logics and ensures communication with the other layers. *Presentation layer* means web browser software for interactive user access. This three-layer architecture is suitable primarily for easy extendibility. Changes in application logics or database scheme can be done without the necessity to intervene in the Presentation layer, and therefore without intervention in the client part of the application.

Then the application can be written in PHP script programming language [4] using the datalogger PHP functions downloaded from [1] and hosted on the server. More gateways can be used by a single application. The application need not be hosted on the user's server, but it is possible to use the server provided within the IQRF platform for this purpose.

Database layer

As a result of analysis, the database scheme has been designed. Then scripts in the SQL language have been generated to create the database. As an actual database server, the **MySQL** [4] has been selected. It is a powerful multiplatform database under GNU General Public License which can be easily implemented for Linux, MS Windows and other operating systems. Communication is possible via the SQL language. Like other SQL databases it is a dialect of this language with several extensions.

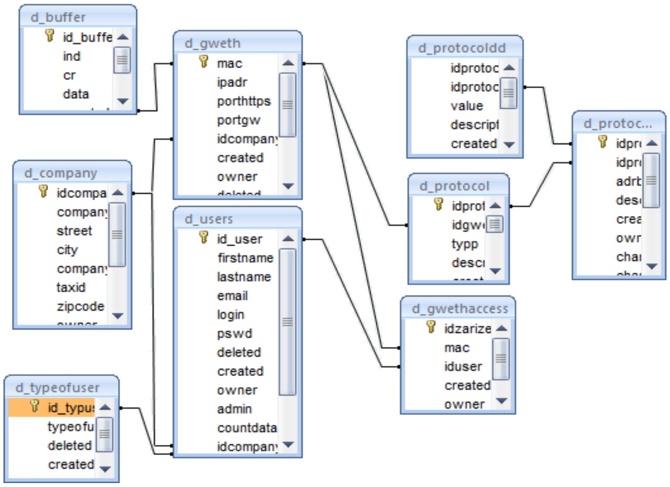


Figure 4. Basic database schema

Application layer

User interface and an interface for graphic outputs from the database are integral parts of the database application. For more comfortable user access to data the web interface has been selected. It is written in PHP.

IV. REAL APPLICATION – SIGMA

One of the real IQRF applications is control system SIGMA for turning of pump vanes. It allows control from the control panel at the pump as well as remotely from the control room. Additional requests are a possibility to control the pump, i.e., to change vane angle or the pump configuration remotely via Internet and supervision over events related to given pump.

To enable also manual control from the GW, GW-QVGA-01 gateway with touch screen can be used instead of GW-ETH-01.

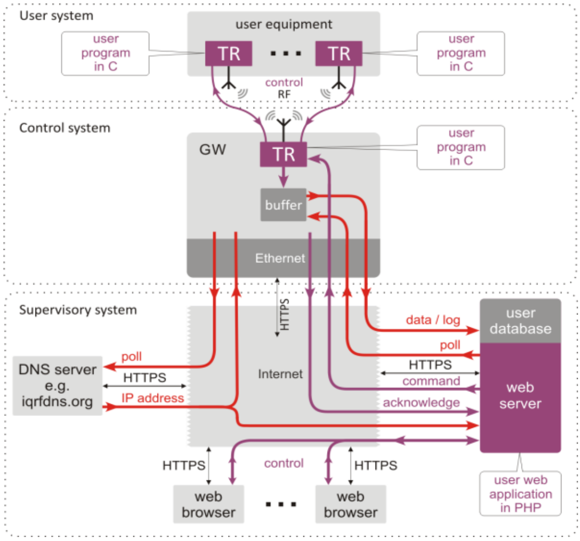


Figure 4. Block diagram of complete system

A. *Analysis*

Requirements:

- Automatic reading from the data buffer
- Data displaying, filtering and search
- Pump configuration
- Changing the angle of pump vanes
- Remote management – status, setup GW-ETH-01 parameters, …

Based on these needs with respect to GW-ETH-01 restrictions followed from parameters mentioned above, a solution using a web application has been developed.

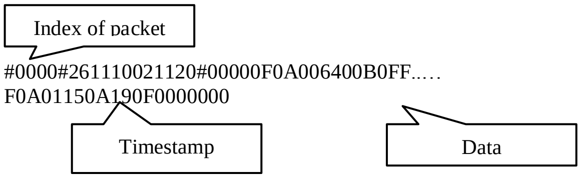
Structure of records (packet) in a data buffer:

A record consists of three parts:

- *index of packet* (2B),
- *timestamp* (6B)
- *data* (41B).

The index of packet range is 0000 – FFFE.

Example of a packet:



#0000#261110021120#00000F0A006400B0FF..…
F0A01150A190F0000000

Maximum number of packets is **146** in the data buffer.

$$\max = \frac{Buffersize}{length(index) + length(timestamp) + length(data)} \quad (1)$$

$$= \frac{7168}{2+6+41} = 146{,}286$$

The buffer is circular (if full, the oldest data is overwritten). Thus, timely buffer reading and data processing must be ensured regularly.

B. *Solutions*

The application includes three parts:

- Automatic reading from the buffer
- Administration part
- User part

1) *Automatic reading from the buffer*

The principle of automatic reading of the buffer:

GW-ETH-01 requests IQRF DNS server to detect its own IP address regularly (in 10 min increments in this case). The IQRF DNS server verifies the MAC (*Media Access Control*) address whether it is in allowed range (00.1F.D5.xx.xx.xx). Then it returns the IP address of a given device. Additionally, the server checks whether the device is registered in application SIGMA. If it is, the server requests GW-ETH-01 to send the data buffer and the GW returns the content. Application SIGMA recognizes new records (compared with last reading) and stores them in the database. This principle is depicted on Fig. 4.

Example of communication:
Request: https://www.iqrfdns.org?IP= 001FD5000048
Answer: Requested IP Address: 81.25.21.74

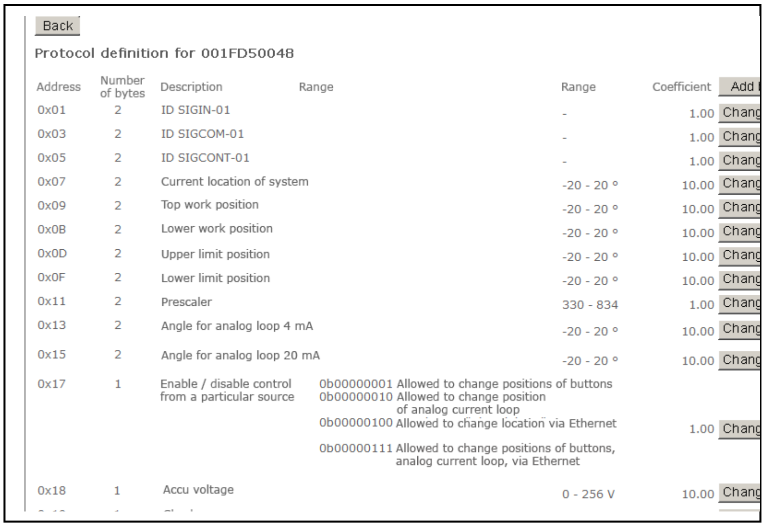
2) *Administration part*

The administration part is intended for:

- Managing of individual registered devices (GW-ETH-01), i.e. adding or removing devices, setup of rights and changing of registration data like the IP address, MAC address, etc.
- User access management of the application, i.e. adding or removing users, changing their rights, usernames, passwords, contact information, etc.
- Defining of individual packets (events) transferred in the network. Based on this definition, the data is displayed then in the user part of the application in a format understandable to humans. Only packets undefined in the network are displayed in HEX code.

The principle of event definitions for individual packets: Visualization of the data of the packet should correspond to user definition. The first byte is the Event header specifying the type of the event. The user can fully define descriptions for individual types to be displayed in the user part of the application. For example, information, error or warning can additionally be differentiated here. The body of the event with variable length follows. Data is visualized in five fields in this example (see Fig. 6): hexadecimal address of the byte in the string, data length, description of the byte, range of allowed values and a coefficient for recalculation.

Figure 6 shows the resulting definition of the event.



Protocol definition for 001FD50048

Address	Number of bytes	Description	Range	Range	Coefficient	Add
0x01	2	ID SIGIN-01		-	1.00	Chang
0x03	2	ID SIGCOM-01		-	1.00	Chang
0x05	2	ID SIGCONT-01		-	1.00	Chang
0x07	2	Current location of system		-20 - 20 °	10.00	Chang
0x09	2	Top work position		-20 - 20 °	10.00	Chang
0x0B	2	Lower work position		-20 - 20 °	10.00	Chang
0x0D	2	Upper limit position		-20 - 20 °	10.00	Chang
0x0F	2	Lower limit position		-20 - 20 °	10.00	Chang
0x11	2	Prescaler		330 - 834	1.00	Chang
0x13	2	Angle for analog loop 4 mA		-20 - 20 °	10.00	Chang
0x15	2	Angle for analog loop 20 mA		-20 - 20 °	10.00	Chang
0x17	1	Enable / disable control from a particular source	0b00000001 Allowed to change positions of buttons 0b00000010 Allowed to change position of analog current loop 0b00000100 Allowed to change location via Ethernet 0b00000111 Allowed to change positions of buttons, analog current loop, via Ethernet		1.00	Chang
0x18	1	Accu voltage		0 - 256 V	10.00	Chang

Figure 6. Example – the RESET event definition

Example of data log:

```
#0000#261110021120#000201040306051400640OCEFF7800B0FFA401C
EFF64000088
```

The value 0x00 (red) indicates a reset event in this particular case.

Example of definition of current angle (see Fig. 6):

It is a 2B information at address 0x07. Recalculation coefficient is 10. Allowed range is from -200 to 200 which means angle from -20.0° .... 20.0°. E.g. 1400, LSB first at this address (see the log example above) means hexadecimal value 0014, decimal value 20 and the angle is 2.0°.

The application is fully generic. Individual definitions of bytes and events can be easily added, modified or deleted.

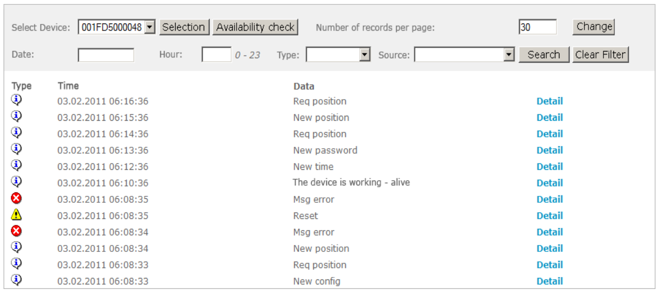
Example: It is discovered during operation that there is a need to include additional information to the packet, e.g. number of revolutions per minute. The administrator just modifies the packet definition by adding the information about the address of given record in the packet, data length and adds its meaning in "human" speech.

An example of data visualization is shown on Figure 7.

Packets can be defined not only for SIGMA, but for various applications. The only restriction is packet-oriented communication. Packets can have different lengths.

User management allows adding and removing users, change data and setup rights. There are three types of users:

- User – with access to User part of application only
- Administrator – with access to both User and Administration parts but with restrictions
- Super administrator – with access to all parts of application without restrictions



Number of records per page: 30

Type	Time	Data	
ⓘ	03.02.2011 06:16:36	Req position	Detail
ⓘ	03.02.2011 06:15:36	New position	Detail
ⓘ	03.02.2011 06:14:36	Req position	Detail
ⓘ	03.02.2011 06:13:36	New password	Detail
ⓘ	03.02.2011 06:12:36	New time	Detail
ⓘ	03.02.2011 06:10:36	The device is working - alive	Detail
✖	03.02.2011 06:08:35	Msg error	Detail
⚠	03.02.2011 06:08:35	Reset	Detail
✖	03.02.2011 06:08:34	Msg error	Detail
ⓘ	03.02.2011 06:08:34	New position	Detail
ⓘ	03.02.2011 06:08:33	Req position	Detail
ⓘ	03.02.2011 06:08:33	New config	Detail

Figure 7. Event list example

3) *User part*

This part is intended for users to handle data, e.g., visualization, filtering and search and if having proper rights also to change pump configuration, e.g., limits and work values. If commands received via Ethernet are allowed, vane angles can be changed too.

Data can be exported to XLS format for further processing (if the user has permission to export).

All activities accomplished in both Administration and User parts are logged. The log is accessible in the Administration part.

V. ACKNOWLEDGEMENT

This research has been supported and co-financed by the Ministry of Industry and Trade of the Czech Republic under contract FR-TI3/254, project "Open Platform for Telemetry".

VI. CONCLUSION

The web application described in this article is designed to be as generic as possible and work with various applications utilizing the IQRF platform. This approach has the following advantages: Visualization is independent of the application layer. Thanks to the IQRF DNS server to read data from the device buffer it is not necessary to program any additional functionality. Handling is very intuitive and no extended training of users is needed. IQRF protocol, the Apache web server, as well as the MySQL database are free resulting in no additional costs required to run such applications.

Specific functionality can be achieved by programming TR in C language using OS functions and web application in PHP script language using the set of PHP functions.

Tests have been completed, and, at present, the SIGMA application is delivered to its final user for operation in practice.

REFERENCES

[1] IQRF:GW-ETH-01 [online]. [retrieved: 3, 2012] http://www.iqrf.org/gw-eth-01.

[2] V. Sulc, R. Kuchta and R. Vrba, “IQRF Smart House - A Case Study", 3rd International Conference on Advances in Mesh Networks, Venice, Italy, 2010

[3] GW-ETH-01 Datalogger User's Guide [online] [retrieved: 3, 2012] http://www.iqrf.org/d=137.

[4] www.mysql.com [retrieved: 3, 2012]

[5] www.php.net [retrieved: 3, 2012]

[6] www.iqrfdns.org [retrieved: 3, 2012]

[7] Raghu Ramakrishnan and Johannes Gehrke: Database Management Systems (Hardcover), McGraw-Hill Science/Engineering/Math; 3 edition (August 14, 2002)

[8] http://www.zigbee.org/ [retrieved: 3, 2012]

[9] http://www.microchip.com [retrieved: 3, 2012]

[10] Vladimir Sulc: 10 reasons WHY IQRF modules are the best choice for IQ applications, RFM Conference, Malaga, Spain, 2004

Web Service-based Applications for Electronic Labor Markets: A Multi-dimensional Price VCG Auction with Individual Utilities

Ricardo Buettner, Jürgen Landes
FOM Hochschule für Oekonomie & Management, University of Applied Sciences
Chair of Information Systems, Organizational Behavior and Human Resource Management
Arnulfstrasse 30, 80335 Munich, Germany
ricardo.buettner@fom.de, juergen.landes@fom.de

***Abstract*—We design an efficient and transaction cost reducing Vickrey-Clarke-Groves auction as part of a web service for the work allocation problem in temporary employment agencies. In this auction bids are work contracts with multi-dimensional salaries. To compute the allocation we assume that every temporary employment worker conveys a utility function specifying the utility gained from working a given job for a salary consisting of multiple components. We then embed the designed mechanism in an updated transaction phase model describing the repeated allocation of temporary agency workers to work assignments. We prove that the designed auction mechanism at the heart of the web service satisfies Incentive Compatibility and Pareto Efficiency.**

***Keywords*-Vickrey-Clarke-Groves auction; web service; electronic human resource management; mechanism design; multi-dimensional price**

I. Introduction

The last decades saw a swift and fundamental change of work and working environments. The number of blue-collar workers has dramatically fallen while the number of white-collar workers has simultaneously increased. This change was driven by the so-called “3-sector-hypothesis” or “Petty’s law” [1]. As a result, many aspects of the working environment became less rigid and numerous employment models have evolved. One of the most successful novel models is that of temporal employment with about 9.5 million employees and a market of more than US$ 340 billion worldwide [2, pp. 11].

Competition pressure creates a sustained impetus for businesses to lower labor costs, which can be achieved in two ways. These costs can be lowered by paying lower wages and salaries or by reducing superfluous transaction costs [3]–[5]. Taking the first route leads to disappointed and unmotivated employees [6]–[8]; thus, we will here focus on the second way instead. Electronic markets are an adequate and a well-established option to reduce transaction costs [9]. During the last two decades electronic markets for commodities were thoroughly investigated and practical business applications (e.g., eBay and Amazon) flourished. Today such electronic markets often incorporate web services [10]–[13], which have also attracted scientific interest on a fundamental level [14].

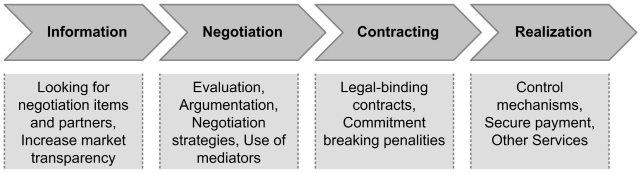

Information
Negotiation
Contracting
Realization
Looking for negotiation items and partners, Increase market transparency
Evaluation, Argumentation, Negotiation strategies, Use of mediators
Legal-binding contracts, Commitment breaking penalities
Control mechanisms, Secure payment, Other Services

Figure 1. Transaction phases according to [18]

By contrast, widespread highly automated electronic labor markets failed so far to materialize. First, unlike well standardized commodities, labor can only imperfectly be described [15, pp. 365]. This imperfect description refers to the description of job demands, to employee characteristics and to job performance. The difficulty in describing job performance due to the complexity of the person-situation interaction has been thoroughly investigated; cf. [16], [17]. Electronic markets for only imperfectly describable goods and services are scarce in the real world; possibly partly due to the little scientific interest they received [18]. Thus e-business applications and web services for such markets have received little attention in the literature.

Second, the utility (function) of work is very complex [19, p. 85] and varies individually. An automated labor market would require market players to specify a priori utility functions (or a similar encoding of personal preferences) specifying preferences for an overwhelming number of possibilities enabling agents to act (search and negotiate) on their behalf.

As a result, today we either find well described theoretical formal models, which are not quite applicable to real world situations, see further Section II, or we find matching algorithms aiding the search for a new job or a new employee [20]. These matching algorithms all address the information phase of a transaction, further transaction phases cannot be supported by such algorithms, see Figure 1. A further automation of electronic labor markets should also address other transaction phases. One such important step would be the development of an efficient allocation algorithm that also computes salaries based on individual and private preferences of market players.

The main contribution of this paper is a web service based application running a novel algorithm (based on a Vickrey Clarke Groves auction [21]–[23]) that matches workers and employers efficiently and that computes salaries consisting of multiple components. In more detail, we consider a Temporary Employment Agency (TEA) that employs Temporary Agency Workers (TAWs) and in turn lends them to businesses for a period of time. We develop a Vickrey Clarke and Groves (VCG) mechanism that allows every TAW and every business representative to specify multi-dimensional utility functions. So, in this auction bids and payments are multi-dimensional. To the best of the authors knowledge such a mechanism has never been described in the literature before.

A. Extended Goal Statement

Summing up we want to develop a web service based application, which uses a novel allocation mechanism, thereby covering at least the first two transaction phases. The aim is that this system satisfies several objectives:

1) Reduce transaction costs.
2) Allocate TAWs to businesses in a Pareto Efficient way, i.e., there is no other way to make no bidder worse off and one better off.
3) Ensure bidders bid their true valuations, i.e., the auction is Incentive Compatible.
4) Enable TAWs to influence their work environment, thereby increasing job engagement and job satisfaction and as a result create added value for businesses [24].
5) Create a work environment that is perceived to be fairer by stakeholders and the general public. Thus improving the social standing of temporary agency workers and temporary work as a whole.

The rest of the paper is organized as follows: Next, we consider related work and the real-world economic background. Then, we present the auction model, followed by an evaluation via mathematical proofs for Incentive Compatibility and Pareto Efficiency of the auction and we give a simple example. Finally, we conclude with a discussion of the model, its limitations and an outlook concerning future work.

II. Real-World Economic Background and Related Work

We now turn to discussing related work.

A. State of the Art in Electronic Negotiations and E-HRM

Being forecast about three decades ago in [25], electronic negotiations have been a hot topic in computer science, so much so that now many well researched overviews exist [26]–[30]. However, the maximal level of automation attainable is controversial. We here exhibit the classification of approaches given in [31].

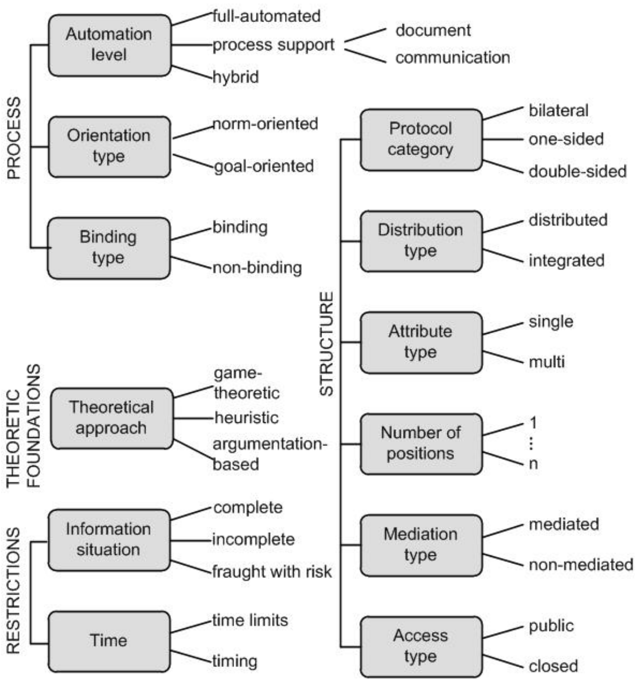


Figure 2. Classification of electronic negotiations [31]

The landscape of scientific research on electronic negotiations is mainly populated by studies of the process and the structure of negotiations whereas issues located in the lower left in Figure 2 have received considerably less attention [31].

The models of electronic negotiations can be classified as game-theoretic, heuristic or argumentation-based [32]. The game-theoretic approach investigates optimal strategies via the analysis of the equilibrium conditions dating back to the seminal work of Nash [33]. Game-theoretic models are well studied, often allowing mathematically elegant investigations, but their potential in practical applications suffers from the assumptions of perfect rationality, unlimited resources and perfect information [32], [34]. Heuristic approaches reject the assumption of unlimited (computing) resources and/or perfect rationality and rather employ thumb rules, see for example [35]. Automated negotiation models based on heuristic approaches have to be intensively evaluated, normally via simulations and/or empirical analysis [32, p. 210]. In argumentation-based negotiations (ABN), agents have the ability to reason their positions. When the negotiation partner is persuaded, who will change her negotiation position, exemplary we mention the system PERSUADER [36].

Related, but not directly relevant, are ongoing developments in e-recruiting and e-HRM, which have been recently surveyed respectively in [37, pp. 231-232] and [38]–[41].

B. Imperfect Information about the Negotiation Item

In [18], 96 electronic negotiation models were studied. Almost all negotiation models (94 percent) assumed imperfect information about the negotiation partner(s). Research on imperfectly described environments and/or negotiation items is considerably less frequent. Only 8 percent of models studied considered the problem of imperfect information concerning the negotiation item, e.g., [42], [43]. Interestingly, one of these models was developed for eHRM [44] and later extended to FuzzyMAN in [45]. In FuzzyMAN and the model implemented therein [43] agents' preferences regarding the negotiation item are expressed in fuzzy [46] terms.

Further work dealing with imperfect information about the negotiation item exists. Different approaches have based their models on different formalisms: probability [47], [48], conjoint scheme [48], genetic algorithms [47] and bandwidths [49].

Overall, we want to develop a game-theoretic model allowing a well-founded evaluation via mathematical proofs. Any successful real-world implementation of such a model has to be comprehensible to all stakeholders [50]. Due to the complex challenges posed by negotiation items that can only imperfectly be described, we use a well-established negotiation model of low complexity, i.e., an auction.

C. Auctions

Auctions are one of the oldest (according to ancient Greek Herodotus, auctions date back to the Babylonians around 500 B.C.) and on the surface simplest form of negotiation. Today, auctions are the main mean to sell expensive antiques, U.S. treasury bonds and rights to use the electromagnetic spectrum for telecommunication purposes. Furthermore, numerous commodity markets rely on auctions [Tsukiji fish market (Tokyo, Japan), the Bloemenveiling flower auction (Aalsmeer, The Netherlands)].

Over the centuries, many auction formats have evolved (first price, second price, open, sealed-bid, with deadline, without fixed deadline, etc.). Different formats were designed to satisfy a variety of properties such as revenue maximization, incentive compatibility and efficiency maximization. Further auction formats were developed, which allow the sale of multiple items at the same time, while other formats discourage collusion and snapping.

More generally, an auction can be understood as a mechanism, which takes as input a set of preferences and outputs an allocation of resources. The art of ensuring that the outcome has desirable properties is known as Mechanism Design (MD) [34]. One branch of MD investigates the design of auctions [51]–[53] to allocate resources to bidders in exchange for a payment.

A Vickrey auction is a sealed-bid second price auction. That is, the auction item is allocated to the highest bidder, who pays the second highest bid submitted. Such an auction satisfies Individual Rationality, Pareto Efficiency and Incentive Compatibility. A Vickrey Clarke Groves (VCG) auction [21]–[23] extends a Vickrey auction allowing multiple items to be auctioned off simultaneously by a single bid taker. Crucially, a VCG auction also satisfies these three properties. Even so, VCG auctions are not always an appropriate mechanism, see for instance [54], [55] and for an overview [56].

Multi-dimensional extensions of classical auctions have been studied. This multi-dimensionality either refers to private valuations (or signals thereof) [57] or to the auction item [58]–[62]. It is well known that in case a public multi-attribute function is used by all participants of an English auction, then such a multi-dimensional auction is equivalent to a one-dimensional auction.

A VCG auction with multi-dimensional bids was developed in [63] by the authors of this paper. In [63] we assumed that all bids were evaluated with respect to the multi-dimensional utility function of the center (TEA). We here build on this work by designing a VCG auction with multiple bidders and multiple bid takers, which all have their own multi-dimensional utility function. The allocation and the payment rule only depend on these functions, in particular they are independent of the TEA's utility function. Results reported in [27] suggest that multi-dimensional auctions yield more utility for the bid taker.

How far, or even if, game theoretic results regarding multi-dimensional prices, respectively multi-dimensional auctions, are transferable to the real world has been investigated in [64], [65].

D. The Business Model of Temporary Employment Agencies

Temporary employment agencies have become a large-scale form of labor market intermediary, acquiring the status of a broker of flexibility at both the micro- and the macro-level [66]. They meet the needs of enterprises to flexibly increase or decrease the size of their workforce, while ensuring for their workers considerable security in terms of job opportunities and employment standards, including pay, working time and training [2, pp. 7, pp. 26]. The business model can be characterized by a triangle. In one corner is the TEA, which has a labor contract with a TAW. Crucially this contract contains a clause granting the TEA the managerial authority to order a TAW to work at (and under the supervision) of one of its clients.

Furthermore, the TEA has business to business (b2b) contracts with customers specifying commercial details of the temporal assignment of TAWs. In general, a TAW working at (and under the supervision of) a client of a TEA and this client do not enter a contract. For the above auction we can hence assume that every participating TAW has a valid work contract with the TEA. Applying the transaction phase model displayed in Figure 1 to the model we developed here, we now adapt the realization phase, see Figure 3. Overall, this yields an adapted model of transaction phases depicted in Figure 4.

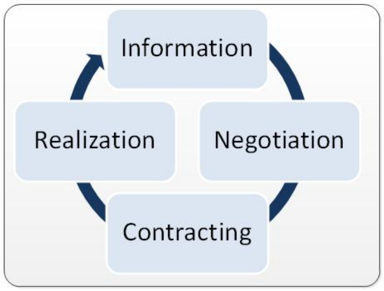


Figure 3. The adapted realization phase

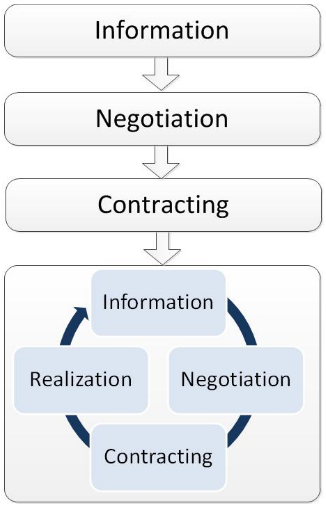


Figure 4. The new model of transaction phases

One important reason for businesses to use a TEA as an intermediary is that the assignment of TAWs may be of limited or unspecified duration with no guarantee of continuation allowing a flexible management of the workforce to adapt to quickly changing market conditions.

Providing a service the TEA charges a fee, ultimately paid by its customers and/or the TAWs. This fee (normally a fixed percentage of the salary) can amount up to 80% of the net salary (depending on circumstances and national laws, e.g., taxes and social security contributions) of a TAW. A considerable part of the operating costs of a TEA are generated by the labor intensive (and hence costly) process of matching TAWs to requests for labor. A further disadvantage shared by TAWs and customers of a TEA is that the matching of TAWs and requests for labor is done to best suit the TEAs' needs. Having no influence over their work environment (including salary) TAWs have in general a lower job engagement, which correlates strongly with productivity [24], [67]. Furthermore TAWs are typically paid less than permanent workers doing the same job violating the principle of same pay for same work causing tensions in the workforce of the client of the TEA [68].

III. The Auction

We now introduce the auction mechanism.

A. Participants

There are two types of participants. First, any business that seeks to hire temporary staff from the TEA can take part. Second, all currently idle TAWs that have a work contract with the TEA may take part. Unless otherwise stated, we mean from now by TAW a participating TAW. A representative of a participating business is from now simply called *employer*. Do note that the TAWs are employed by the TEA. The term employer is chosen here to ease the understanding and the write-up; see further Section II-D. To ease the notation, we make the convention that every employer is looking to fill exactly one full time vacancy (multiple vacancies at a company are modeled by multiple employers).

B. Information Phase

During the information phase participants search for potential matches, read background information stored at the TEA or on the Internet on potential employers (policies, corporate philosophy and identity) or on TAWs (CV, references and possibly a sample of previous work). To predict future potential job performance of job candidates (TAWs), employers may carry out e-assessments [69], [70] enhanced by exchanged emails, interviews conducted via text-based chat applications and/or (video) calls. Similarly, TAWs may pick up information crucial for their valuation of future job assignments. From a formal perspective, the sending and exchanging of signals, indices and arguments can be seen to take place to combat the infamous adverse selection problem [15], [71].

To enhance quality and speed of the search in large databases, a recommender system [72], [73] and/or a reputation system [74], [75] may be used.

C. Bidding

Let $E := \{E_1, \ldots, E_e\}$ be the set of employers seeking to secure the services of a TAW and let $W := \{w_1, \ldots, w_t\}$ be the set of TAWs looking for work. For $1 \leq i \leq n$ let M_i be a salary component, such as wage per hour, benefits, sick pay or overtime premiums. Let $M = M_1 \times \ldots \times M_n$ be the set of all contracts consisting of these components. We define an additive structure on M by $\oplus M \times M \to M$ via addition by component $(m_1, \ldots, m_n) \oplus (k_1, \ldots, k_n) := (m_1 + k_1, \ldots, m_n + k_n)$.

For $1 \leq i \leq t$ let $\{E_{i_1}, \ldots, E_{i_{k(i)}}\}$ be the set of employers interested in acquiring the services of TAW w_i. Now every $w_i \in W$ sends a utility function $u^i_{i_r} : M \to \mathbb{R}$ to employer E_{i_r} detailing how much s/he (dis-)likes to work for E_{i_r}, if the TAW is sufficiently qualified to perform these jobs. These utility functions are also communicated to the TEA and to all other employers. In case TAW w_i does not send a utility function to an employer E_{i_r}, the utility function $u^i_{i_r}$ is set to be the zero function. We can hence assume that the TAWs' utility functions are functions mapping $M \times E \to \mathbb{R}$.

A significant proportion of TAWs is low-skilled [2, table 3.6 page 19] and might hence require training and/or decision support tools to construct these utility functions; see further [76] for one such tool designed for an electronic labor market. These tools normally use preference elicitation techniques [77]. Such techniques can go a long way to aid the understanding and thus acceptance of designed applications and computer systems [50], [78].

We assume that every employer E_d is risk neutral, fully rational and the valuation of TAWs and contracts is independent of the valuation of other employers. We hence assume that E_d has a utility function $UU_d : M \times W \to \mathbb{R}$ specifying how much value a TAW working a certain job for a given contract brings to employer E_d. We assume furthermore that these functions are *additive*, that is UU_d is given as a sum of utility functions, i.e., $UU_d(m, w) = U_d(m) + V_d(w)$. This notion of an additive utility function generalizes the notion of a quasi-linear utility function to multi-dimensional prices. Furthermore, we assume that U_d commutes with the additional structure $\oplus$.

At this point, every employer has a choice to make based on the utility functions $u^i_{i_r}$ communicated, either to take part in the following auction and accept the binding outcome or to drop out and not take part in the auction. For the time being, we assume, for the sake of a simpler notation, that no employer drops out. Why an employer might drop out will be investigated in the section Incentives.

Every employer now makes one sealed bid for each TAW, from which a not vanishing utility function was received. That is for $1 \leq d \leq e$ a, in general partial, function $bid_d : W \to M$ is communicated to the TEA. These functions are in general partial because not every $w \in W$ communicates a utility function to all employers.

To ease the notation we make the following convention. Every partial function bid_d is extended to a total function by setting bid_d to zero wherever it was not defined. Furthermore $u^i_\emptyset$ and $bid_\emptyset$ are set to vanish everywhere.

D. The Allocation

Definition 1 A function $f : \{1, \ldots, t\} \to \{1, \ldots, e\} \cup \{\emptyset\}$ is called an *allocation* if and only if $f(i) = f(k)$ implies that $f(i) = f(k) = \{\emptyset\}$. Thus an allocation allocates every employer (representing a single vacancy) at most one TAW.

The TEA then calculates the allocation f that maximizes

$$\sum_{1 \leq l \leq t} u^l_{f(l)}(bid_{f(l)}(l)) \tag{1}$$

under the constraint that for $f(l) \neq \emptyset$ employer $E_{f(l)}$ has put in a non-zero bid for TAW w_l. The constraint implies that an employer will never be allocated a TAW w_l, for which this employer has not put in a bid.

For $1 \leq d \leq e$ let f_d be the allocation, which maximizes the sum in Equation 1 and which satisfies the constraint in case that E_d does not enter a single bid (or equivalently E_d does not take part in the auction).

All participants are then informed of the outcome of the allocation concerning themselves. So every employer E_d learns, which TAW (if any) has been allocated to work for E_d, vice versa for the TAWs. To calculate the salary (in auction terminology: payment rule) we need to introduce some notation.

E. Salaries

Definition 2 For $t \in \mathbb{N}$ let $[t] := \{1, \ldots, t\}$ and for $1 \leq l \leq t$ put $[t - l] := \{1, \ldots, t\} \setminus \{l\}$ and $[t - \emptyset] := [t]$. Let $g : X \to Y$ be a function, then the *level set of g at level y* is defined as $\{x \in X | g(x) = y\}$.

For $l \in [t], x_l \in \mathbb{R}$ and a utility function $U_d : M \to \mathbb{R}$ let $\langle \sum_{l \in [t]} x_l \rangle_d$ be an element in M that minimizes $U_d(\oplus_{l \in [t]} m_l) = \sum_{l \in [t]} U_d(m_l)$ under the condition that for every $l \in [t]$ m_l is an element of the level set of $u^l_{f_d(l)}$ at level x_l. That is employer E_d gets to pick an element in all those level sets. Since this expression will later be part of a salary paid, the employer makes choices suiting best his/her needs. For $x \in \mathbb{R}, m \in M$ and a utility function $U_d : M \to \mathbb{R}$ let $x - U_d(m)$ be an element of the level set of U_d at level $x - U_d(m)$.

Employer E_d then pays TAW $w_{f^{-1}(d)}$

$$\begin{aligned} Salary(E_d) := & - \sum_{l \in [t - f^{-1}(d)]} u_f(l)^l(bid_{f(l)}(l)) \\ & + \langle \sum_{l \in [t]} u^l_{f_d(l)}(bid_{f_d(l)}(l)) \rangle_d. \end{aligned} \tag{2}$$

Do note that the second term in Equation 2 cannot be influenced by any bids made by employer E_d, since it only contains terms that are calculated for an auction, in which she did not participate. To ease the reading we set $U_d(C_d)$ to be the utility received from this term.

Note that employer E_d wants to maximize the overall utility received, which equals

$$V_d(f^{-1}(d)) + \sum_{l \in [w - f^{-1}(d)]} u^l_{f(l)}(bid_{f(l)}(l)) - U_d(C_d). \tag{3}$$

Any fully rational bidding strategy an employer pursues will hence only depend on the first term in 2 and the TAW allocated due to our assumptions about UU_d (additive and commuting with addition).

F. Incentives

Theorem 1 The above auction satisfies Incentive Compatibility and Pareto Efficiency.

Proof: The main idea in the following proof is to show the fact that it is in every employers best own interest to maximize the utility to be distributed. That is, a rational selfish employer seeks to pursue the common good.

Firstly, we have to prove, that bidding their true valuation is an ex-post Nash equilibrium for all bidders. That is, even knowing all other bids, it is for every bidder an optimal strategy to bid true values. We here mean by true valuation that bid_d satisfies $V_d(l) = u_d^l(bid_d(l))$ for all $l \in [t]$. So the employer obtains as much utility from being allocated w_l as the bid for w_l by this employer is worth to w_l.

Recall that f maximizes the sum in Equation 1. Now if $V_d(l) = u_d^l(bid_d(l))$ for all $l \in [t]$, then Equation 1 and Equation 3 only differ by a constant. So f also maximizes 3 in this case. There is hence no better strategy for employer E_d than to bid the private true valuations for all TAWs.

Let us now assume for the second part of the proof that all bidders bid their true valuations (i.e., they all follow an optimal strategy). Then f maximizes the overall utility distributed. Hence, allocating more utility to one bidder will at least make one other bidder lose utility. ■

In case there is less than full confidence in the TEA to properly execute the auction and/or to keep information entered into the system private, an auction issuer [79] can be used to ensure the proper handling of sensitive information and to ensure the correct computation of the allocation and payments.

Example 1 Consider an auction with three employers $\{E_1, E_2, E_3\}$, which have decided to bid for a TAW w_1. Assume furthermore that for the three utility functions communicated to the employers it holds that $u_i = u$. If $u_1(bid_1(1)) > u_2(bid_2(1)) > u_3(bid_3(1))$, then w_1 will work for employer E_1 for a salary in the level set of u at level $u(bid_2(1))$ to be specified by E_1.

From this example, the following observation can be inferred. If there is only one auction item (i.e., one TAW) and the TAW is only interested in the salary (i.e., not in the jobs to do), then the winning bidders bid is of lower utility (to the bidder and the TAW) than the salary paid.

Do note that the above calculations were all done without the explicit knowledge of U_d, that is the private valuation of E_d of multi-dimensional salaries. To actually calculate the figure in Equation 2 one needs to know U_d. In one-dimensional price VCG-auctions all U_d are simply assumed to be the identity function $id : \mathbb{R} \to \mathbb{R}$ and furthermore it is assumed that this is *public* knowledge. It is hence not surprising, alas not ideal, that the here presented mechanism cannot do without any knowledge of the U_d. Assuming that the U_d are known to the TEA or assuming a certain knowledge of the level sets of the U_d are two ways of solving this problem (it suffices to know one element in every level set of the U_d and the level sets containing the C_d).

Observe that in the one-dimensional case the level sets completely determine the function. Counterintuitively, this multi-dimensional price auction requires less information about utility functions on prices than the one-dimensional counterpart.

By contrast, note that the bidder's utility from obtaining an auction item (i.e., a TAW) is revealed through the design of the mechanism, if the bidder acts rationally.

Finally we have to consider the case of an employer E_d that is not allocated a TAW. To keep the attractive properties of Incentive Compatibility and Pareto Efficiency, the payment rule also has to be applied to such an employer. A payment goes to or comes from the center (i.e., the TEA), as there was no worker allocated to this employer. The payment can be calculated and subsequently paid in case one of the above two conditions on the knowledge of the U_d is satisfied. Note that in case every employer is allocated a TAW, this issue concerning the payment rule does not surface.

As we have seen above it makes sense for bidders to be honest but what about the TAWs? Recall that they also submitted utility functions; can they obtain an advantage by not reporting their true valuations? We have already seen in the above example that misreporting the shape of the utility functions $u_{i_r}^i$ is in general not advantageous.

Recall that for bidders it is rational to bid such that $V_d(l) = u_d^l(bid_d(l))$. So a TAW stands to gain by making extraordinary high demands. To discourage such behavior the $u_{i_r}^i$ are communicated to all bidders, which have subsequently the option to abstain from the auction in case salary demands are perceived to be too high. An employer not participating in the auction will look elsewhere for workers.

G. The Algorithmic Complexity of Calculating the Allocation

The number of complete matchings in a connected complete bipartite graph with independent sets of sizes $x \geq y$ is $\frac{x!}{(x-y)!}$. So the number of possible allocations with $e \geq t$ is $\frac{e!}{(e-t)!}$. Hence calculating the allocation f that maximizes the utility is of high algorithmic complexity [80].

Observe that the problem of calculating this allocation simplifies significantly in case the bipartite graph consists of several disconnected components. Connected components of a bipartite graph can be found in linear time. From a practical point of view, reducing the problem to connected components of the graph is hence highly desirable. If the computational complexity of calculating the allocation after the decomposition into connected components is still too high for practical purposes, then approximation algorithms [81], [82] can be used to calculate an allocation that is close to the efficient allocation.

H. The Aftermath

Consider a TAW allocated to a given employer and recall that the VCG-mechanism outputs work contracts consisting of multiple salary components. Possibly there is a contract, which both the TAW and the employer prefer to the one generated by the mechanism. This is in stark contrast to the one-dimensional case with only one salary component. There an employer prefers a lower and a TAW a higher salary. Allowing renegotiations of multi-dimensional salaries may yield gains for the TAW and the employer (and possibly the TEA); however, it renders the above mechanism Incentive Incompatible.

IV. Possible Extensions and Further Applications

Do note that we assumed above that every TAW can only have one job at the same time. This is surely a sensible assumption if all jobs are full time jobs. Extending the above auction to also include part time jobs is possible; one then has to use multivalued allocations f (instead of functions) that assign TAWs to (possibly) multiple employers. Again, this new mechanism does satisfy Incentive Compatibility and Pareto Efficiency. Due to space constraints and our wish not to overload this paper with notation we will refrain here from doing so.

Furthermore, it is possible to include externalities in the mechanism by allowing for the possibility that the utility functions u_l^d depend not only on the job w_l will be working but on the whole allocation f. For example this enables a TAW to express that s/he prefers to work at the same place as her/his husband/wife, yielding monetary gains (lower transportation costs by using the same car) and non-monetary gains (joint lunch). From a formal point of view, extending the framework in this way does not yield; in our opinion; valuable insights and we will hence not present it here.

Conversely the framework can be extended to allow bidders to bid for multiple TAWs simultaneously instead of single TAWs. So the operator of a restaurant can put in a combined bid for a cook and a waiter, which have previously successfully worked together, which may be higher than the sum of bids for the cook and waiter individually.

The here presented mechanism can of course be also used to allocate tasks in other circumstances, for instance in grid and cloud computing similar allocation problems need to be solved. The tasks to be allocated are computing tasks. One further area of application is the wide field of social choice dealing with the multi-faceted problem of how to increase social welfare [83].

V. Conclusion and Future Work

We have presented a web service based application running an algorithm matching TAWs and business, which uses a multi-dimensional price VCG auction. In this auction, the TAWs can individually express salary demands, depending on the job to be done and the employer. We showed that the best a bidder can do is to bid true valuations. Furthermore, we have seen that there are also incentives for the TAWs to honestly report preferences. We have hence designed a mechanism encouraging proper behavior creating an environment that hopefully contributes to a rise in the social standing of temporary workers and temporary work in general.

Formally, we have applied a model of transaction phases to our approach and subsequently extended this model, see Figure 4. This new model allows us to state that our approach addresses the online information and negotiation transaction phase inside the realization phase thus allowing a further automation of a particular labor market. We are optimistic that electronic auctions are a suitable mean to reduce transaction costs for trading goods and services that cannot perfectly be described, in particular labor. Enabling market players (here TAWs and employers) to specify their own multi-dimensional utility functions is in our view a key ingredient for a successful implementation.

Overall, we have reached the goal we set out [see Section I-A] and alleviated in the last section highlighted drawbacks of the business model of a TEA.

A. Limitations

The here presented approach is limited by the assumption that all salary components can be added in a natural way, furthermore we assumed that the employers utility functions are additive and commute with addition. A restriction to the numbers of participants taking part in the auction arises from the complexity of calculating the allocation f. Furthermore, the assumptions of full rationality and risk neutrality are in general not always satisfied in the real world.

One limiting factor in electronic labor markets is the human aversion to new technologies. However, an easy-to-use, understandable and benefiting system stands good chances to mostly overcome such aversions [78].

B. Future Research

In our view, it is desirable to design a mechanism similar to the above that can handle salary components, which cannot be added canonically (such as: job title, job location, work task). We consider the long-term goal of a development and an implementation of a multi agent system (with agents acting for and on behalf of market players) for electronic labor markets worthy of future attention from the scientific community as well as from business communities.

Acknowledgment

The authors gratefully acknowledge support from grant 17103X10 from the German federal ministry of education and research.

References

[1] Y. Murata, "Engel's law, Petty's law, and agglomeration," *J Dev Econ*, vol. 87, no. 1, pp. 161–177, 2008.

[2] International Labor Organization - WPEAC2009. (2009) Private employment agencies, temporary agency workers and their contribution to the labour market. International Labor Organization. WPEAC/2009. [Online]. Available: http://www.ilo.org/wcmsp5/groups/public/---ed_norm/---relconf/documents/meetingdocument/wcms_122432.pdf

[3] R. H. Coase, "The Nature of the Firm," *Economica*, vol. 4, no. 16, pp. 386–405, November 1937.

[4] ——, "The Problem of Social Cost," *J Law Econ*, vol. 3, no. 1, pp. 1–44, October 1960.

[5] O. Williamson, *The economic institutions of capitalism*. The Free Press, 1985.

[6] C. E. Jurgensen, "Job preferences (What makes a job good or bad?)," *J Appl Psychol*, vol. 63, no. 3, pp. 267–276, 1978.

[7] W. A. Kahn, "Psychological Conditions of Personal Engagement and Disengagement at Work," *Acad Manage J*, vol. 33, no. 4, pp. 692–724, December 1990.

[8] D. M. Rousseau, "New hire perceptions of their own and their employer's obligations: A study of psychological contracts," *J Organ Behav*, vol. 11, no. 5, pp. 389–400, 1990.

[9] T. W. Malone, J. Yates, and R. I. Benjamin, "Electronic markets and electronic hierarchies," *Commun. ACM*, vol. 30, no. 6, pp. 484–497, 1987.

[10] H. Demirkan, R. J. Kauffman, J. A. Vayghan, H.-G. Fill, D. Karagiannis, and P. P. Maglio, "Service-oriented technology and management: Perspectives on research and practice for the coming decade," *Electron Commer Res Appl*, vol. 7, no. 4, pp. 35 –376, 2008.

[11] S. Dustdar and W. Schreiner, "A survey on web services composition," *Int J Web Grid Serv*, vol. 1, no. 1, pp. 1–30, 2005.

[12] M. P. Papazoglou, P. Traverso, S. Dustdar, and F. Leymann, "Service-Oriented Computing: State of the Art and Research Challenges," *IEEE Computer*, vol. 40, no. 11, pp. 38–45, 2007.

[13] M. P. Papazoglou, *Web Services: Principles and Technology*. Addison-Wesley, 2008.

[14] P. Maglio and J. Spohrer, "Fundamentals of service science," *J Acad Market Sci*, vol. 36, pp. 18–20, 2008.

[15] M. Spence, "Job market signaling," *Q J Econ*, vol. 87, no. 3, pp. 3555–374, 1973.

[16] J. Hackman, "Toward understanding the role of tasks in behavioral research," *Acta Psychol (Amst)*, vol. 31, pp. 97–128, 1969.

[17] T. A. Judge, D. Heller, and M. K. Mount, "Five-factor model of personality and job satisfaction: A meta-analysis," *J Appl Psychol*, vol. 87, no. 3, pp. 530–541, 2002.

[18] R. Buettner, "Electronic Negotiations of the Transactional Costs Perspective," in *Proceedings of WWW/Internet*, vol. 2. IADIS, 2007, pp. 99–105.

[19] S. Kraus, "Negotiation and cooperation in multi-agent environments," *Artif Intell*, vol. 94, no. 1-2, pp. 79–97, 1997.

[20] W. Gates and M. Nissen, "Designing agent-based electronic employment markets," *Electron Commerce Res*, vol. 1, no. 3, pp. 239–263, 2001.

[21] W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J Financ*, vol. 16, no. 1, pp. 8–37, 1961.

[22] E. H. Clarke, "Multipart pricing of public goods," *Public Choice*, vol. 11, no. 1, pp. 17–33, 1971.

[23] T. Groves, "Incentives in teams," *Econometrica*, vol. 41, no. 4, pp. 617–631, 1973.

[24] J. A. Gruman and A. M. Saks, "Performance management and employee engagement," *Hum Resour Manage R*, vol. 21, no. 2, pp. 123–136, 2011.

[25] R. Davis and R. G. Smith, "Negotiation as a metaphor for distributed problem solving," *Artif Intell*, vol. 20, no. 1, pp. 63–109, 1983.

[26] G. Anandalingam, R. W. Day, and S. Raghavan, "The Landscape of Electronic Market Design," *Manage Sci*, vol. 51, no. 3, pp. 316–327, 2005.

[27] M. Bichler, G. Kersten, and S. Strecker, "Towards a Structured Design of Electronic Negotiations," *Group Decis Negot*, vol. 12, pp. 311–335, 2003.

[28] A. R. Lomuscio, M. Wooldridge, and N. R. Jennings, "A Classification Scheme for Negotiation in Electronic Commerce," *Group Decis Negot*, vol. 12, pp. 31–56, 2003.

[29] F. Lopes, M. Wooldridge, and A. Novais, "Negotiation among autonomous computational agents: principles, analysis and challenges," *Artif Intell Rev*, vol. 29, pp. 1–44, 2008.

[30] M. Ströbel and C. Weinhardt, "The Montreal Taxonomy for Electronic Negotiations," *Group Decis Negot*, vol. 12, pp. 143–164, 2003.

[31] R. Buettner, "A Classification Structure for Automated Negotiations," in *Proceedings of WI-IATW*. IEEE, 2006, pp. 523–530.

[32] N. Jennings, P. Faratin, A. Lomuscio, S. Parsons, M. Wooldridge, and C. Sierra, "Automated Negotiation: Prospects, Methods and Challenges," *Group Decis Negot*, vol. 10, pp. 199–215, 2001.

[33] J. F. Nash, "The Bargaining Problem," *Econometrica*, vol. 18, no. 2, pp. 155–162, 1950.

[34] R. K. Dash, N. R. Jennings, and D. C. Parkes, "Computational-Mechanism Design: A Call to Arms," *IEEE Intell Syst*, vol. 18, pp. 40–47, 2003.

[35] S. Kraus, *Strategic negotiation in multiagent environments*. The MIT Press, 2001.

[36] E. P. Sycara, "Resolving adversarial conflicts: an approach integration case-based and analytic methods," Ph.D. dissertation, Georgia Institute of Technology, 1987.

[37] I. Lee, “Modeling the benefit of e-recruiting process integration,” *Decis Support Syst*, vol. 51, no. 1, pp. 230–239, April 2011.

[38] S. Lang, S. Laumer, C. Maier, and A. Eckhardt, “Drivers, challenges and consequences of E-recruiting: a literature review,” in *Proceedings of SIGMIS-CPR*. ACM, 2011, pp. 26–35.

[39] S. D. Maurer and D. P. Cook, “Using company web sites to e-recruit qualified applicants: A job marketing based review of theory-based research,” *Comput Human Behav*, vol. 27, no. 1, pp. 106–117, 2011.

[40] S. Strohmeier, “Research in e-HRM: Review and implications,” *Hum Resour Management R*, vol. 17, no. 1, pp. 19–37, 2007.

[41] ——, “Concepts of e-HRM consequences: a categorisation, review and suggestion,” *Int J Hum Resour Man*, vol. 20, no. 3, pp. 528–543, 2009.

[42] X. Luo, N. R. Jennings, N. Shadbolt, H. Leung, and J. H. Lee, “A fuzzy constraint based model for bilateral, multi-issue negotiations in semi-competitive environments,” *Artif Intell*, vol. 148, no. 1-2, pp. 53–102, 2003.

[43] F. Teuteberg, “Experimental evaluation of a model for multilateral negotiation with fuzzy preferences on an agent-based marketplace,” *Elctron Markets*, vol. 13, no. 1, pp. 21–32, 2003.

[44] K. Kurbel and I. Loutchko, “Multi-agent Negotiation under Time Constraints on an Agent-based Marketplace for Personnel Acquisition,” in *Proceedings of MALCEB*, 2002, pp. 566–579.

[45] F. Teuteberg and I. Loutchko, “FuzzyMan: An Agent-based E-Marketplace with a Voice and Mobile User Interface,” in *Software Agent-Based Applications, Platforms and Development Kits*, R. Unland, M. Klusch, and M. Calisti, Eds. Birkhäuser, 2005, ch. 5, pp. 281–306.

[46] L. A. Zadeh, “Fuzzy sets,” *Inform Comput*, vol. 8, no. 3, pp. 338–353, 1965.

[47] A. Cardon, T. Galinho, and J.-P. Vacher, “Genetic algorithms using multi-objectives in a multi-agent system,” *Rob Auton Syst*, vol. 33, no. 2-3, pp. 179–190, 2000.

[48] G. E. Kersten and S. J. Noronha, “WWW-based negotiation support: design, implementation, and use,” *Decis Support Syst*, vol. 25, no. 2, pp. 135–154, 1999.

[49] S. Matwin, S. Szpakowicz, Z. Koperczak, G. E. Kersten, and W. Michalowski, “Negoplan: An Expert System Shell for Negotiation Support,” *IEEE Expert*, vol. 4, pp. 50–62, December 1989.

[50] B. Dineen, R. Noe, and C. Wang, “Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences,” *Hum Resour Manage*, vol. 43, no. 2-3, pp. 127–145, 2004.

[51] P. Klemperer, “Auction Theory: A Guide to the Literature,” *J Econ Surv*, vol. 13, no. 3, pp. 227–286, 1999.

[52] ——, “What really matters in auction design,” *J Econ Perspect*, vol. 16, no. 1, pp. 169–189, 2002.

[53] V. Krishna, *Auction Theory*. Academic Press, 2002.

[54] M. Rothkopf, T. Teisberg, and E. Kahn, “Why are Vickrey auctions rare?” *J Polit Econ*, vol. 98, no. 1, pp. 94–109, 1990.

[55] M. Jackson, “Non-existence of equilibrium in Vickrey, second-price, and English auctions,” *Rev Econ Des*, vol. 13, pp. 137–145, 2009.

[56] T. Sandholm, “Issues in computational Vickrey auctions,” *Int J Electron Comm*, vol. 4, pp. 107–129, March 2000.

[57] D. Levin, J. Peck, and L. Ye, “Bad news can be good news: Early dropouts in an English auction with multi-dimensional signals,” *Econ Lett*, vol. 95, no. 3, pp. 462–467, 2007.

[58] F. Branco, “The design of multidimensional auctions,” *Rand J Econ*, vol. 28, no. 1, pp. 63–81, 1997.

[59] Y. Che, “Design competition through multidimensional auctions,” *Rand Journal of Economics*, vol. 24, no. 4, pp. 668–680, 1993.

[60] Y. De Smet, “A multicriteria perspective on reverse auctions,” *4OR Q J Oper Res*, vol. 5, pp. 169–172, 2007.

[61] J. E. Teich, H. Wallenius, J. Wallenius, and A. Zaitsev, “A multi-attribute e-auction mechanism for procurement: Theoretical foundations,” *Eur J Oper Res*, vol. 175, no. 1, pp. 90–100, 2006.

[62] S. E. Thiel, “Multidimensional auctions,” *Economics Letters*, vol. 28, no. 1, pp. 37–40, 1988.

[63] J. Landes and R. Buettner, “Job allocation in a temporary employment agency via multi-dimensional price VCG auctions using a multi agent system,” in *CPS Proceedings of MICAI2011*. IEEE, 2011, pp. 182–187.

[64] H. Estelami, “Consumer perceptions of multi-dimensional prices,” *Adv Consum Res*, vol. 24, pp. 392–399, 1997.

[65] A. Lange and A. Ratan, “Multi-dimensional reference-dependent preferences in sealed-bid auctions - How (most) laboratory experiments differ from the field,” *Games Econ Behav*, vol. 68, no. 2, pp. 634–645, 2010.

[66] N. M. Coe, J. Johns, and K. Ward, “Mapping the Globalization of the Temporary Staffing Industry,” *Prof Geogr*, vol. 59, no. 4, pp. 503–520, 2007.

[67] W. H. Macey, B. Schneider, K. M. Barbera, and S. A. Young, *Employee engagement: Tools for analysis, practice, and competitive advantage*. Wiley-Blackwell, 2009.

[68] J. S. Adams, “Toward an understanding of Inequity,” *J Abnorm Soc Psychol*, vol. 67, no. 5, pp. 422–436, November 1963.

[69] D. Bartram, “Testing on the Internet: Issues, Challenges and Opportunities in the Field of Occupational Assessment,” in *Computer-Based Testing and the Internet*, R. K. H. Dave Bartram, Ed. John Wiley & Sons, Ltd, 2008, ch. 1, pp. 13–37.

[70] S. Laumer, A. von Stetten, A. Eckhardt, and T. Weitzel, “Online Gaming to Apply for Jobs - The Impact of Self- and E-Assessment on Staff Recruitment,” in *Proceedings of HICSS*, 2009, pp. 1–10.

[71] G. A. Akerlof, “The Market for ’Lemons’: Quality Uncertainty and the Market Mechanism,” *Q J Econ*, vol. 84, no. 3, pp. 488–500, 1970.

[72] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions,” *IEEE T Knowl Data En*, vol. 17, pp. 734–749, 2005.

[73] J. B. Schafer, J. Konstan, and J. Riedi, “Recommender Systems in E-Commerce,” in *Proceedings of E-commerce*, 1999, pp. 158–166.

[74] K. Hoffman, D. Zage, and C. Nita-Rotaru, “A Survey of Attack and Defense Techniques for Reputation Systems,” *ACM Comput. Surv.*, vol. 42, pp. 1–31, 2009.

[75] 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.

[76] A.-F. Rutkowski and B. Van De Walle, “Cultural Dimensions and Prototypical Criteria for Multi-Criteria Decision Support in Electronic Markets: A Comparative Analysis of Two Job Markets,” *Group Decis Negot*, vol. 14, pp. 285–306, 2005.

[77] L. J. Savage, “Elicitation of personal probabilities and expectations,” *J Am Stat Assoc*, vol. 66, no. 336, pp. 783–801, 1971.

[78] F. D. Davis, “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology,” *Mis Quart*, vol. 13, no. 3, pp. 319–340, September 1989.

[79] M. Naor, B. Pinkas, and R. Sumner, “Privacy preserving auctions and mechanism design,” in *Proceedings of E-commerce*. ACM, 1999, pp. 129–139.

[80] C. H. Papadimitriou, *Computational complexity*. John Wiley and Sons Ltd., 2003.

[81] N. Nisan and A. Ronen, “Computationally feasible VCG mechanisms,” *J Artif Intell Res*, vol. 29, no. 1, pp. 19–47, 2007.

[82] T. Sandholm, “Algorithm for optimal winner determination in combinatorial auctions,” *Artif Intell*, vol. 135, no. 1-2, pp. 1–54, 2002.

[83] K. J. Arrow, *Social Choice and Individual Values*, 2nd ed. Yale University Press, 1963.

Cross-Domain Query Navigation System for Touchscreens by Exploiting Social Search History

Ryo Shimaoka
Faculty of Policy Management Studies
Keio University
5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
s09411rs@sfc.keio.ac.jp

Shuichi Kurabayashi
Faculty of Environment and Information
Keio University
5322 Endo, Fujisawa, Kanagawa 252-0882, Japan
kurabaya@sfc.keio.ac.jp

***Abstract*—Tablets and smartphones have gained immense popularity in recent times, and it is envisaged that they will increasingly be the devices of choice for users accessing the Internet. However, the user interface of conventional Web search engines, which employ keywords that require many taps by the user, are unsuitable for mobile terminals, which are normally equipped with touchscreens. We propose a cross-domain query navigation system that reduces the taps required for inputting queries by providing a content-dependent *word map* that presents the relevance between keywords. This *word map* presents keywords that enable both narrowing action, whereby users append a new keyword to specify the context of a query, and sliding action, whereby users replace a keyword to change the query context. The *word map* is unique in that it recommends queries for narrowing and sliding transitions by computing these two types of directional relevance between input keyword and another keyword in the log. The system is applicable to the existing query logs of search engines, social networking services, and users' browsers, enabling users to control the term recommendation by selecting the logs to be analyzed. The recommendation may be a commonly recognized relevant term from the global query logs of search engines or a personalized term from the user's browser history.**

Keywords-Query Navigation; Personalization; User Interface; Collective Intelligence; Web Search Engine.

I. Introduction

The recent years have witnessed a rapid rise in the popularity of tablet devices and smartphones, and concomitantly, a widespread increase in the use of the touch-based user interface (UI). Statistics published by Cisco indicate that the global mobile data traffic grew 2.6-fold in 2010, nearly tripling for the third consecutive year [1]. In addition, statistics published by Google Confidential and Proprietary suggest that by 2015, more than a quarter of the mobile traffic will be used for information retrieval and that the number of Internet users not using PC devices will increase to 788 million [2]. Hence, a major shift in use of Internet-connected devices, from PCs to mobile terminals, is currently underway.

A large portion of Internet activity is in the form of queries to search engines. However, many users have difficulty querying a search engine on a complex topic that encompasses several terms, such as "JavaScript and HTML5" or "ActionScript and API," relating to a subject with which they are not familiar. Mobile devices present an additional difficulty: although touchscreens are generally very convenient for other functions, they are not very convenient to use as a typing tool. In particular, queries in Chinese-Japanese-Korean-Vietnamese (CJKV) languages present special difficulties because each CJKV character requires two or three input strokes. In mobile devices, predictive input methods are the predominant method for supporting the input of long sentences and terms. These predictive input methods recommend terms and sentences that can be concatenated to the user's input character sequence. Another conventional method is a keyword suggestion approach, such as Google Suggest. When a user inputs an initial query term, this method suggests related terms by calculating inter-term relevance, exploiting the search engine's query log to recognize the relevant terms.

However, these conventional methods are based on the co-occurrence probability and hence are unsuitable for inputting queries that consists of several cross-domain terms, such as "climbing health care costs." In such cases, predictive input methods may not correctly recommend the next search term, and a cross-domain term-relevance calculation is required. On the other hand, Google Suggest will not tailor the search results to an individual user's interests, because it uses standardized search terms drawn from a universal users' log. Thus, the UIs provided by a conventional Web search engine require users to tap many times, making these UIs unsuitable for mobile terminals.

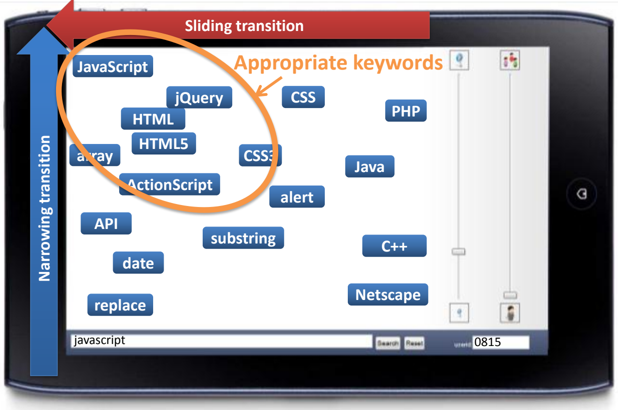


Figure 1. User Interface of a Cross-Domain Query Navigation System.

This paper proposes a cross-domain query navigation system that assists in the input of multiple queries by show

ISBN: 978-1-61208-200-4

a content-dependent *word map* to present the relevance between keywords. This system allows users to input keywords by selecting an appropriate keyword in a convenient manner, because the word map shows the next coordinate instantly, as shown in Figure 1. Here, we explain an example scenario of query navigation shown in Figure 2. This figure shows the following two types of navigation:

● Narrowing navigation: Users append a new keyword (e.g., "traffic," or "global") to specify the context of a query. The appended keyword is at a lower level of abstraction than those of the existing keywords.

● Sliding navigation: Users replace a keyword to change the context of a query. Here, a user removes an existing keyword (e.g., "traffic") that is not within the scope of the current topic of interest and inserts another one (e.g., "user") that is relevant to the current topic of interest, thus shifting the focus of the query. In this case, the system recommends a new keyword (e.g., "laptop") appropriate in the current context.

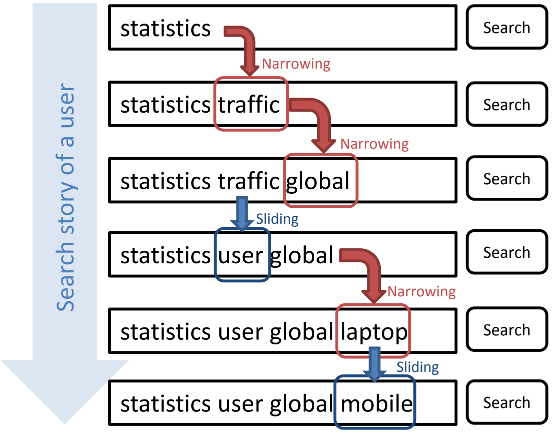


Figure 2. Narrowing and Sliding Transitions in Query Construction.

The advantage of this system is that it obviates the need for users to enter the subsequent search terms themselves; instead, they are able to select from among those that are mapped on the screen. This search story approach makes it possible to reduce taps. For example, when a user wishes to add the search terms "global," "mobile," and traffic" to the term "statistics," which has already been inserted in the search box, only one tap is required for each term, making three in all. The keywords are presented after considering the user's browser history, which enables personalization, and the other users' querying history, which supports a user by exploiting collective intelligence. Our system configures the balancing between personalization and collective intelligence support dynamically, that is not possible with conventional search engines.

Another advantage of the system is afforded by the fact that it is also applicable to the search stories of social networks, which include groups of experts in various fields, as shown in Figure 3. This allows users to search within a domain that they are not familiar with, by drawing on the collective knowledge and experience of expert groups through their search stories. Furthermore, the application would also help users construct a query in a language that they do not know well.

II. Related Work

The query expansion method is a well-known means of helping a search engine's users to input complex queries. The traditional example of query expansions is Google Suggest, which recommends keywords from a uniform set that is derived from all users based on the number of previous searches. Currently, many researchers are focusing on personalization mechanisms in query expansion [3]. For example, Teevan, et al. [4] proposes a personalization method that considers the user's specific interests by constructing a user profile from the relevance feedback in a ranking. Gauch, et al. [5] proposes an implicit personalization mechanism that generates ontology-based user profiles without user feedback, by monitoring the user's browsing activities.

An alternative method of query expansion uses the concept of *community*. Smyth, et al. [6] introduces the collaborative filtering method, which exploits a similar relationship between queries and result pages for each community. The method expands a query by referring to a graded mapping between users and items.

The most significant difference between our approach and the conventional ones listed above is the concept of *query dimension*. Our system provides two *dimensions* in the query building process: narrowing and sliding. Narrowing is a typical query building method that allows users to increase the specificity of a query after starting with an initial keyword. Our approach also supports sliding, which suggests cross-domain keywords by computing the implicit relevance of keywords in different domains, such as "climbing," "health," "care," and "costs." To increase the precision of the sliding process, our approach exploits the search story of a relevant group or community.

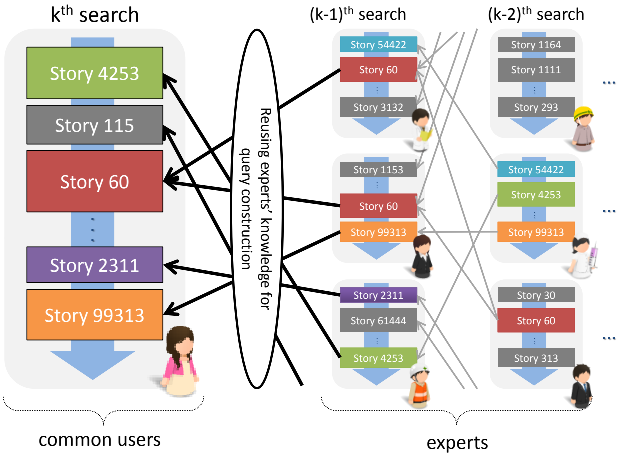


Figure 3. Search Story Sharing among Users Empowers the System's Cross-Domain Keyword Recommendation.

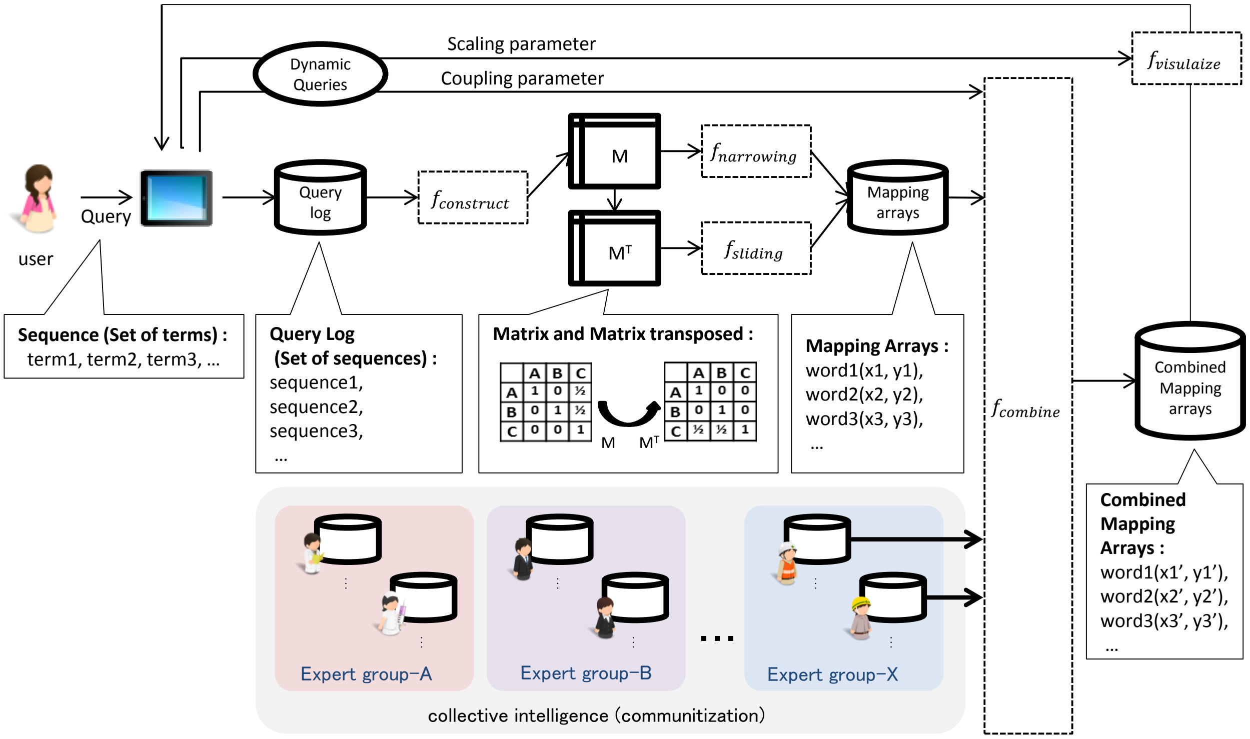


Figure 4. System Architecture of Recommendation of Search Terms on Word Map for Directional Relevance between Input Keyword and Another Keyword in the Log.

III. Approach

The narrowing and sliding forms of navigation are based on an *inter-term relationship matrix* constructed from a query log, as shown in Figure 4; the purpose of this matrix is to record the relationship between keywords for each user. The system converts the matrix into recommendation scores, which correspond to the coordinate values for narrowing and sliding as presented on the user interface. The system combines the recommendation scores from the user and from social groups within the domain of interest.

The first stage is for the system to construct the matrix from the query log, a set of keyword sequences recorded when the user inputs a complex query in a search box. This matrix contains scores representing relationships between search terms. It is updated from the query log. In the second stage, the system converts the matrix into two types of recommendation scores: sliding and narrowing. The system calculates these scores based on computation of the inner product of the matrix and the matrix transpose. The final stage is to combine the recommendation scores of the user with that of social groups within the domain of interest. Our concept of social-network-based relevance computation is reusing third parties' knowledge about query construction. This system may distinguish several groups of users by using other SNS's social graph, such as Twitter's follower/followee structure and Facebook's friend structure. The user can also adjust the parameters of the combination process.

A. Data Structure

The data structure in this system consists of two data elements—*query log* and *inter-term relationship matrix*—which are now explained in detail as follows.

1) Query Log

A query log is a set of sequences that consist of search terms. We define *Log (L)* as a data structure that is determined based on a *Sequence (S)* of keywords inputted by a user. L_i of user *i* is defined by the following equation.

$$L_i := \langle S_0, S_1, \cdots, S_n \rangle$$

where *n* is the number of sequences.

A sequence is a set of searched keywords. Therefore, we define a *Sequence (S)* as a data structure that is determined based on the *keyword (k)*. S_j is defined by the following equation.

$$S_j := \langle k_0, k_1, \cdots, k_n \rangle$$

where *n* is the number of keywords.

2) Inter-Term Relationship Matrix

A function generates a relationship matrix from the query log. The relationship matrix contains a set of values that represent the directional relevance between each pair of keywords (the *weight* of the association). This is a square matrix, whose rows and columns each correspond to the same set of keywords. We define the *Matrix (M)* of user i as a data structure that is determined based on the *weight (w)*.

$$M_i := \begin{bmatrix} w_{[0,0]} & \cdots & w_{[n,0]} \\ \vdots & \ddots & \vdots \\ w_{[0,n]} & \cdots & w_{[n.n]} \end{bmatrix}$$

where n is the number of keywords. The system also generates the matrix transpose ${M_i}^T$ for reverse look-up.

B. Functions

The proposed system provides three main functions. The first function constructs the relationship matrix from a query log. The second function converts the matrix into narrowing and sliding scores for recommendations. The final function combines the recommendation scores of the user with those of a social group can provide expertise concerning the user's domains of interest.

1) Constructing a Matrix from a Query Log

The system provides a fundamental function to construct a matrix from a query log. The function is defined as follows.

$$f_{construct}(L_i) \rightarrow M_i$$

where M_i contains a set of values $w_{[l,\ m]}$ that represent the weight of the directional relevance between k_l and k_m.

This function updates the matrix every time the user inputs a query. Thus, we set the *weight (w)* of a sequence (S_j) as the relevance, determined based on the *rank* of *keyword (k)*.

$$w_{[l,m]}(S_j) \rightarrow \left[\frac{1}{rank(k_0 \in S_j)}, \frac{1}{rank(k_1 \in S_j)}, \cdots, \frac{1}{rank(k_n \in S_j)}\right].$$

Figure 5 shows an example of this summation process.

2) Converting a Matrix into Recommendation Scores

The system provides a fundamental function to convert a matrix into mapping arrays. Each mapping array contains the vertical and horizontal scores of a given keyword in relation to the *origin* keyword, i.e., the last term of a query. Thus, the function f_{map} generates two directional relevance scores, such as sliding relevance and narrowing relevance, according to a keyword specified as the origin point. $f_{map}(o)$ is defined as follows:

$$f_{map}(M_i, o) \rightarrow \{\langle p_v, p_h\rangle_0, \dots, \langle p_v, p_h\rangle_n\},$$

$$\langle p_v, p_h\rangle_k \rightarrow \langle \sum_{j=0}^{n} M_{i[o,j]} \cdot M_{i[j,k]}, \sum_{j=0}^{n} {M_i}^T{}_{[o,j]} \cdot {M_i}^T{}_{[j,k]} \rangle$$

where p_v and p_h are the vertical and horizontal scores, respectively, for the word map. The vertical score corresponds to the directional relevance of a narrowing search, whereas the horizontal score corresponds to the directional relevance of a sliding search.

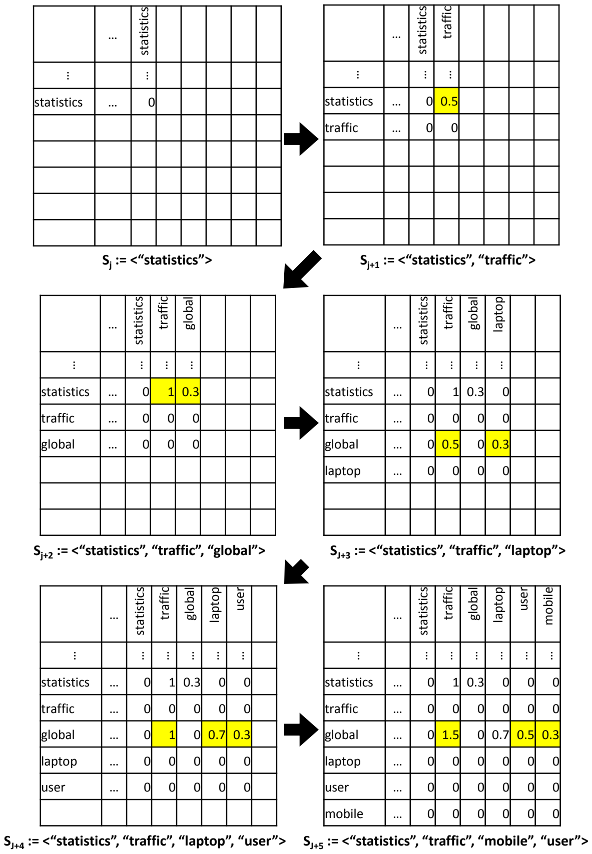
3) Combining the Recommendation Scores of the User and the Expert Groups

This system uses the collective expertise of other users for its recommendations. This recommendation function merges the matrix of a user and the matrices of other search engine users in a weighted combination. The combination weighting, or *rate*, is set by the user of this system via a slider on the Web page. Thus, we define $f_{combine}$ as a function that is determined based on a *combination rate (r)*.

$$f_{combine}(p, G, r) \rightarrow [\frac{p \cdot (100\ -\ \mathrm{r}) + \frac{\sum_{\mathrm{j=0}}^{\mathrm{n}} G}{n} \cdot \mathrm{r}}{100}, \cdots]$$

where p is a correlation score and n is number of persons in *group (G)*.

These equations combine the matrix of the main user with the average of the matrices of all users to yield a final score.



	…	statistics					
⋮		⋮					
statistics	…	0					

S_j := <"statistics">

	…	statistics	traffic				
⋮		⋮	⋮				
statistics	…	0	0.5				
traffic	…	0	0				

S_{j+1} := <"statistics", "traffic">

	…	statistics	traffic	global			
⋮		⋮	⋮	⋮			
statistics	…	0	1	0.3			
traffic	…	0	0	0			
global	…	0	0	0			

S_{j+2} := <"statistics", "traffic", "global">

	…	statistics	traffic	global	laptop		
⋮		⋮	⋮	⋮	⋮		
statistics	…	0	1	0.3	0		
traffic	…	0	0	0	0		
global	…	0	0.5	0	0.3		
laptop	…	0	0	0	0		

S_{J+3} := <"statistics", "traffic", "laptop">

	…	statistics	traffic	global	laptop	user	
⋮		⋮	⋮	⋮	⋮	⋮	
statistics	…	0	1	0.3	0	0	
traffic	…	0	0	0	0	0	
global	…	0	1	0	0.7	0.3	
laptop	…	0	0	0	0	0	
user	…	0	0	0	0	0	

S_{J+4} := <"statistics", "traffic", "laptop", "user">

	…	statistics	traffic	global	laptop	user	mobile
⋮		⋮	⋮	⋮	⋮	⋮	⋮
statistics	…	0	1	0.3	0	0	0
traffic	…	0	0	0	0	0	0
global	…	0	1.5	0	0.7	0.5	0.3
laptop	…	0	0	0	0	0	0
user	…	0	0	0	0	0	0
mobile	…	0	0	0	0	0	0

S_{J+5} := <"statistics", "traffic", "mobile", "user">

Figure 5. Matrix Composition Process.

IV. IMPLEMENTATION

We have implemented a prototype system for evaluating the recommendation of search terms by analyzing users' query logs. This system has been coded in full-stack JavaScript language, which implies that the server-side and client-side modules are implemented in JavaScript only.

1) Modules

The engine of this system has two main modules for the server side and client side. These modules use the same data structure, which is a user's search story, but they serve two different functions. On the server side, the system provides communitization, whereas on the client side, it provides personalization.

The server-side module of the prototype system outputs four types of arrays: narrowing and sliding scores for both the user and the community. The advantage of these outputs is that the system is able to present search terms with just the client-side module. Therefore, this module is only run when the user inputs a new query.

The client-side module presents search terms on the user interface. The system presents the candidate search terms in a two-dimensional space that is defined by the narrowing axis and the sliding axis. The search terms are positioned in relation to parameters that the user inputs using two sliders, one of which defines the combination rate for other search stories (community) and the other defines the scaling rate (zoom factor) for words.

2) User Interface

The user interface of this system consists of the word map, zoom slider, social slider, search box, and search button (Figure 6). The most important control is the social slider. This slider defines the extent to which the user's search story is combined with the community search stories. The system allows users to discover appropriate keywords by adjusting the combination level of search terms, if no terms are found initially.

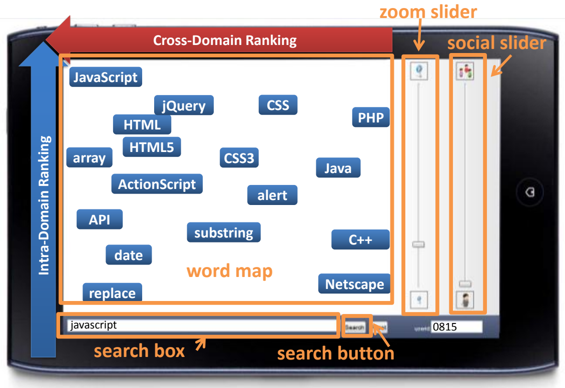


Figure 6. User Interface of Cross-Domain Query Navigation System.

The procedure of this system is as follows:

Step 1: The user inputs the initial keyword of the query in the search box, and the system presents keywords on the word map.

Step 2: The user taps an appropriate keyword. The system displays the keyword in the search box and presents a new set of keywords on the word map (Figure 6 shows only one term in the search box. The system displays the next keyword when the user enters it via the touch interface).

Step 3: If no appropriate keyword is shown, the user may drag the social slider until the combination level generates a satisfactory range of keywords.

Step 4: The user repeats Steps 2 and 3 as necessary. The user taps the search button and the system retrieves the search results.

Figure 7 shows an example of how the word map can be changed using the social slider. The arrows in this figure show how the keywords move when the slider is operated. The origin point (upper left) corresponds to the initial query "JavaScript". The system provides candidate keywords from an expert group of Web designers but not programmers, displaying new candidate keywords that are used by Web designers, such as "sample code," "Web design," and "Flash," but not keywords that are used by programmers, such as "Java" and "C++." The figure shows that the keywords that are more often used by Web designers than by programmers, such as "HTML5," are slightly shifted from lower right to upper left. The figure shows that the keywords that are more often used by programmers than by Web designers, such as "API," are shift from upper left to lower right.

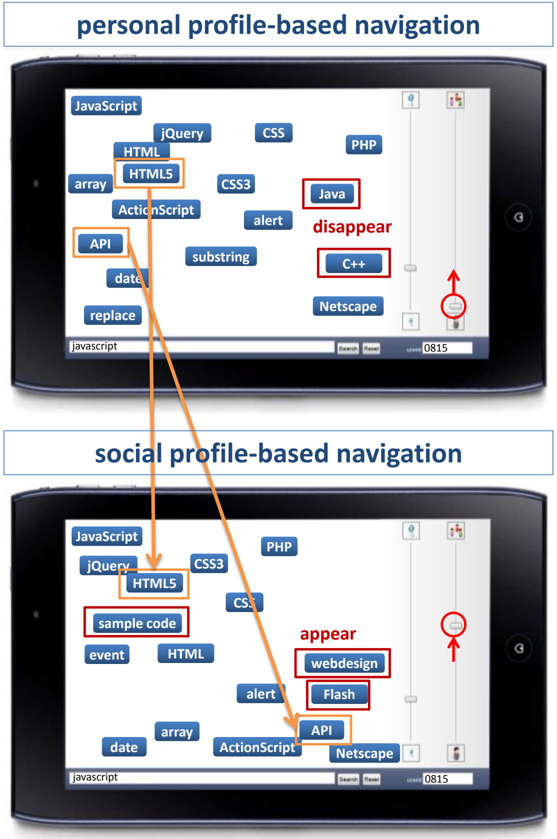


Figure 7. Change in the Keyword Positions on Word Map with Social Slider.

V. Experiments

This section presents experimental studies that clarify the effectiveness of our approach. In particular, this experiment evaluates the f_{map} function that converts a matrix into recommendation scores. The experiment evaluates the directional relevance between the input search terms and the candidate search terms, based on the user and the community query logs. The experiment compares the narrowing, which is the legacy keyword recommendation, and the sliding, which is the original feature of this system. Due to the limitation of the space, this experiment clarifies that our approach calculates the appropriate distance between query keywords by asking 10 test subjects to evaluate the navigation results.

A. Overview

This experiment measures the precision of the keyword ranking of the recommended query keywords by comparing the manually-conducted correct result set and our system. We have set up the inter-term relationship

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

matrix by submitting 952 queries to Google. As a result, we obtained a 108 × 108 matrix. We have designed the three test topics, “design,” “e-book,” and “editorial,” as the initial keywords. These three test cases generate three rankings. We select the top 10 keywords of narrowing / sliding relevancy in each search topics, such as “design,” “e-book,” and “editorial,” as shown in TABLE I.

In order to verify the precision of the results, we had 10 persons evaluate the rankings. 10 test subjects evaluated relevance of 60 keywords and three topics from the viewpoint of *narrowing* and *sliding*. Every subject rated each recommended keyword according to the following 5-point scheme: 0 (completely not relevant), 1 (not relevant), 2 (slightly relevant), 3 (relevant), and 4 (very relevant). We considered the ideal ranking as the average of 10 results.

Table I. NARROWING AND SLIDING KEYWORDS AND THE RANKS.

design			e-book			editorial		
rank	narrowing	sliding	rank	narrowing	sliding	rank	narrowing	sliding
1	editorial	e-book	1	design	editorial	1	color	e-book
2	layout	editorial	2	editorial	research	2	design	design
3	color	research	3	color	implication	3	e-book	research
4	image	history	4	layout	history	4	layout	implication
5	scheme	magazine	5	image	program	5	magazine	hisitory
6	research	book	6	scheme	genre	6	newspaper	program
7	ranking	program	7	research	magazine	7	history	genre
8	magazine	genre	8	ranking	book	8	electronic	magazine
9	newspaper	retrieval	9	search	retrieval	9	television	ranking
10	iPhone	iTV	10	engine	brief	10	iPhone	brief

B. *Evaluation Result*

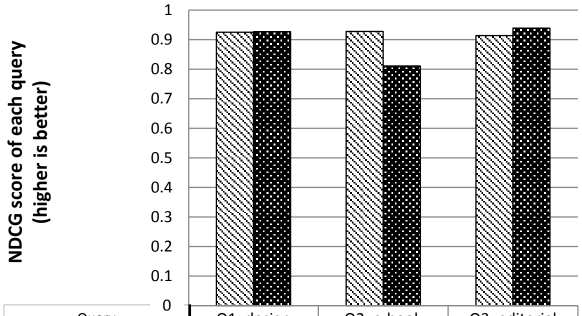
This experiment compares the keyword ranking of narrowing and sliding recommendation with evaluations by test subjects. The evaluation of this experiment applies normalized discounted cumulative gain (NDCG).

$$\mathrm{DCG} = \sum_{\mathrm{i}=1}^{10} \frac{rel_i}{log_2 i},\ \mathrm{IDCG} = \sum_{\mathrm{i}=1}^{10} \frac{rel\prime_i}{log_2 i},\ \mathrm{NDCG} = \frac{DCG}{IDCG}$$

where rel_i is the average evaluation score given by the test subjects, and $\mathrm{rel'}_\mathrm{i}$ is the average scores in descending order. Figure 8 shows NDCG of narrowing and sliding recommendation for three topics (“design,” “e-book,” and “editorial”). Higher score means a better retrieval precision. The most important result is a score of *sliding recommendation* because the narrowing recommendation is close to the conventional query recommendation method. The NDCG of sliding recommendation is the almost same as that of narrowing recommendation. This result implies that our sliding recommendation achieves highly practical precision, although the sliding recommendation generates different keywords from the narrowing recommendation. By using this system, the user received the precise query keyword, which shared a cross-domain relationship with the initial keyword. This recommendation is a very powerful tool to input a complex query consisting of cross-domain keywords.

VI. CONCLUSION AND FUTURE WORKS

We have proposed the complex query navigation system that exploits search stories of social groups. This system recommends candidates for a next search term by calculating the directional relevance along two conceptual dimensions and performing narrowing and sliding operations. A social combination function enables the user to utilize the knowledge of social groups to facilitate navigation. We implemented a prototype system that is able to retrieve and present candidate keywords for multiple queries while reducing the number of taps required. As a future work, we plan to develop a social-based query recommendation mechanism and to evaluate scalability in a complex query navigation in multiple domains.



Query	Q1: design	Q2: e-book	Q3: editorial
▨ narrowing (conventional method: single domain recommendation)	0.925012297	0.927528221	0.913642768
■ sliding (our method: cross-domain recommendation)	0.926944072	0.811268464	0.939018707

Figure 8. NDCG of Narrowing and Sliding Recommendation.

REFERENCES

[1] Cisco. “Global Mobile Data Traffic Forecast Update, 2010-2015” - Cisco Visual Networking Index, February 1, 2011, http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper_c11-520862.pdf . [retrieved: 5, 2012]

[2] Google. “Admod - Tablet Survey,” March 2011, http://www.ccapitalia.net/descarga/docs/2011-AdMob-TabletSurvey.pdf .

[3] A. Micarelli, F. Gasparetti, F. Sciarrone, and S. Gauch, “Personalized search on the World Wide Web,” in The Adaptive Web, LNCS 4321/2007, 2007, vol. 4321, pp. 195-230.

[4] J. Teevan, S. T. Dumais, and E. Horvitz, “Personalizing search via automated analysis of interests and activities,” in Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR ’05, 2005, p. 449.

[5] S. Gauch, J. Chaffee, and A. Pretschner, “Ontology-based personalized search and browsing,” in Web Intelligence and Agent Systems - IOS Press vol. 1, no. 3-4/2003, pp. 219-234.

[6] B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell, “Exploiting query repetition and regularity in an adaptive community-based Web search engine,” User Modeling and User-Adapted Interaction, Apr. 2005, vol. 14, no. 5, pp. 383-423.

Towards Mobile Energy-Adaptive Rich Internet Applications

Johannes Waltsgott
Faculty of Computer Science
Technische Universität Dresden, Germany
johannes.waltsgott@tu-dresden.de

Klaus Meißner
Faculty of Computer Science
Technische Universität Dresden, Germany
klaus.meissner@tu-dresden.de

***Abstract*—Composite web applications built from reusable components are replacing traditional, monolithic Rich Internet Applications (RIAs). Based on the rising number of smartphones and the increasing usage of mobile applications, composite web applications arrive on mobile devices. While the computing power of the latter rapidly grows, an unsolved problem persists: the limited energy resources of mobile devices. We propose an architecture for mobile energy-adaptive RIAs, which allows for energy optimization by adapting the distribution of components between the server and the device and minimizing the data communication.**

***Keywords*-composite applications, mobile applications, energy efficiency, component migration, mashups**

I. Introduction

Service-oriented architectures are current practice to build reusable and agile composite applications from loosely coupled services and resources. Lately, this composition paradigm has been deployed to the presentation layer as well. This paved the way for mashups, or *composite web applications*, as an alternative to former, monolithic RIAs. Mashups have gained acceptance for consumers as well as enterprises [1].

At the same time, the number of smartphones sold has risen tremendously. A Gartner survey shows that, compared to the same period of 2010, smartphone sales increased by 42 % in the third quarter of 2011, with an even higher estimation for Q4 and early 2012 [2]. Not only has the number of devices been growing, but also the usage of mobile applications. Today, emailing, web browsing, personal navigation and social media applications are used by many smartphone users. In the future, new usage scenarios will arise, making excessive use of the device's sensors [3]. In parallel, mobile applications rely heavily on remote data and cloud storage to overcome limitations regarding storage on the device and to support collaborative scenarios. However, there exists one thing, which does not even rudimentarily keep pace with the increasing distribution, performance and usage of smartphones: the device's battery capacity (cf. [4]). Since mobile devices are generally required to last as long as possible, the limited energy budget and severe energy consumers are ongoing issues for smartphone users.

In this paper, we introduce our approach towards mobile energy-adaptive RIAs, in which we capitalize on the mobile communication management at application layer. We focus on composite applications based on CRUISe [5], a universal composition platform for mashups.

The paper is structured as follows. In Section II, we give a brief overview of the CRUISe composition platform. Section III summarizes the challenges and related work regarding energy-aware mobile applications. In Section IV, we introduce our proposal and describe its respective parts. Finally, we discuss our findings and outline future work in Section V.

II. The CRUISe Composition Platform

Our approach towards energy-efficient RIAs is based on the CRUISe Platform for universal mashup composition, whose principles have been introduced in [6]. CRUISe extends the known service-oriented paradigm to include the presentation layer. Applications consist of universal parts, which provide data access, business logic and UI. These CRUISe *components* share a generic component model and a platform-independent description language, the *Semantic Mashup Component Description Language* (SMCDL, see [7] for more details). The inner workings of a component are encapsulated by an interface consisting of three abstract concepts, namely *property*, *operation* and *event* (cf. [8]). The public state of a component is represented by its properties, while changes of the inner state result in publishing events, which could be consumed by the runtime or other components. The functionality provided by a component is accessible by calling its operations. This allows for a loose coupling of components, where an event-based communication architecture routes event messages from publishing components to the respective subscribers via *event channels*.

A composite application is described by a generic *composition model* [8], referencing the involved components' IDs, the required event channels and layout information. A service-oriented infrastructure supports the dynamic execution of CRUISe applications at runtime, as depicted in Figure 1. The composition model (upper left of figure) is interpreted by a CRUISe *runtime environment*, shown in the middle, that brings the composition to life, using universal components provided by the service layer at the bottom.

The runtime environment receives the component code of every component from a *component repository*, shown

ISBN: 978-1-61208-200-4

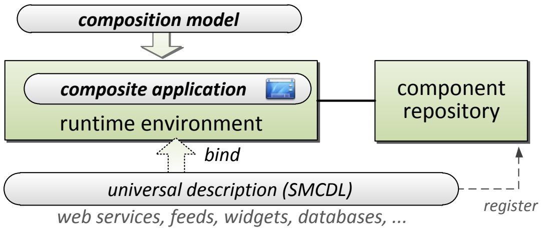


Figure 1. Architectural overview of the CRUISe composition platform [6]

on right, where all components are registered. Finally, the runtime integrates and instantiates the components and establishes the specified event channels.

Thus far, CRUISe provides a mature platform for universal mashup composition, which allows for dynamic component integration at runtime and is already in practical use in industry. However, the existing capabilities of the CRUISe runtime lack support for energy-aware execution of CRUISe applications. This is of high importance, especially when used on mobile devices, as motivated before.

III. Challenges and Related Work

First of all, it is crucial to determine the main energy consumers of current smartphones. Carrol and Heiser performed a detailed analysis of a mobile phone's power consumption [9], identifying the display and graphics as the main consumers, followed by the GSM radio while the CPU, generally, is of lower relevance. They profiled phone components isolated by several benchmarks and measured the energy consumption for various usage scenarios, e. g., audio and video playback, text messaging, phone calls, emailing and web browsing. The results show that the display (LCD panel, touchscreen, graphics and backlight) head the power consumption of all none-GSM-intensive scenarios. Otherwise, e. g., phone calls, emailing and web browsing, the GSM radio respectively the WiFi system consumed the most of power, while WiFi showed a noticeable higher energy efficiency for data transfer. Since the backlight of the display is either automatically dimmed by the operating system or explicitly set to a user-specified value, optimizing a device's communication behavior provides the most promising approach for energy optimization at application level.

A specialized analysis of energy consumption of mobile communication (GSM, 3G and WiFi) was performed by Balasubramanian et al. [10]. Their results show a fair energy efficiency for smaller data size (10 KB) using GSM and for bigger data size (>100 KB) using WiFi. 3G consumes significantly more energy for data transfer, according to the high *tail energy*, which covers the energy consumed while remaining in a high power transmitting state even after the actual data transfer is completed. Based on their measurements, they derived a power model for all three communication technologies covering ramp, transition and tail energy as well as the tail time. Besides, they proposed *TailEnder*, a network protocol to be integrated in mobile operating systems, which schedules data transfer for delay-tolerant applications or prefetches data (e. g., search results) for suitable usage scenarios. It has to be surveyed, whether TailEnder could be used in addition to our approach, since it is a very data-centric view. Besides, its suitability for highly interactive RIAs has to be proven yet.

Based on the understanding of mobile communication as relevant energy consumer, it is of high importance to influence the communication behavior and distribution of mobile applications. MAUI [4] is an approach to optimize the device runtime by energy-aware distribution of mobile code. It relies on special attributes in the code, marking methods to be (potentially) executed on remote hosts. A profiler collects context information on the device, the network and the program state at runtime. A solver processes the information and decides whether a method should be invoked locally or remote, taking the overhead (transfer time and cost, processing cost) into account. Since we focus on component-based RIAs, where the components act as black boxes and are handled by their interface description only, MAUI seems not suitable.

A more coarse-grained approach proposes a method for energy-efficient workflow distribution [11], in which a workflow model is enhanced by a network model and a data model. The network model describes valid environments (mobile or hosted) for a workflow's activity. The data model describes the transmission costs between two activities, which are derived from the data size to be transmitted and the power model presented in [10]. The most efficient distribution of the activities is calculated by a *minimal cut algorithm*, applied to a *cost graph*. Afterwards, the workflow is deployed accordingly. Their evaluation showed average energy savings up to 37 % for optimized distribution. However, since their approach focuses on workflows with determined size of data transmitted between activities, it does not suit highly interactive RIAs. Moreover, we strive for energy optimization at runtime rather than at application deployment.

Flinn proposed remote execution for mobile applications [12], focusing on energy-aware adaptation of application quality to optimize energy consumption by delivering application performance to meet user requirements. However, he did not account for communication costs while distributing application code but concentrated on network bandwidth and latency as performance parameters only.

In summary, it can be stated, that relevant energy consumers in current smartphones, which could be controlled by system or application level, have been identified clearly: the mobile communication devices. Thus, research focuses on optimizing mobile communication: from fine-grain code level to coarse-grain approaches. Nevertheless, there are shortcomings regarding energy optimization for highly interactive, component-based RIAs, whose communication behavior could not easily be predicted prior to runtime.

IV. Proposal for Solution

Facing the afore mentioned drawbacks, we introduce *energy-adaptive Rich Internet Applications* (eRIAs): an architecture and runtime environment for composite mobile web applications, which allows for energy-efficient reconfiguration of applications and migration of components between client and server. It consists of three main parts, as depicted in Figure 2: (1) a *context monitor*, collecting information on the application and device state and on user requirements at runtime; (2) a *migration manager*, deciding, which reconfiguration of a component distribution between server and mobile device is the most energy-efficient one for a given context and migrating the affected components accordingly; (3) a *runtime environment*, supporting the dynamic execution of application components on the server as well as on the client. In the following, we discuss the respective parts of our proposal in detail.

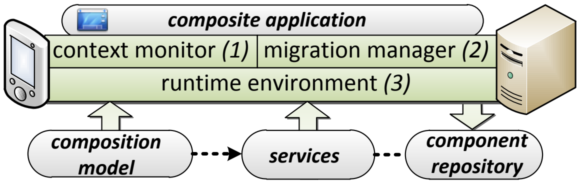


Figure 2. Architectural overview of the eRIA proposal

A. Context Monitor

CRUISe-based composite web applications consists of UI and service components, loosely coupled by an event bus. Components are encapsulated by an interface, describing *properties*, *operations* and *events* by the SMCDL. Thus, their inner state is unknown to the runtime environment, which requires monitoring their behavior externally. The Context Monitor shall collect information on the component's communication traffic to allow for a calculation of the energy cost. Communication between components can be measured at the Event Manager (EM), which is part of the CRUISe runtime environment (cf. [13]), since the EM is responsible for wiring components by delivering event messages to operation calls according to the definition of the composition model. Communication with external services can be measured at the Service Access (SA, also part of the CRUISe runtime environment), which acts as a global proxy for external requests due to client-side security restrictions (cf. [13]). The SA delivers the received data via a given callback method to the component.

Besides, the Context Monitor shall gather information on the energy context of the mobile device. These information involve the current battery state, whether the device is just charging, what kind of mobile communication technology is used and basic parameters for signal strength and bandwidth. Finally, the Context Monitor shall be able to collect user requirements, e. g., a user forces a high-performance mode of an application approving a higher energy consumption knowing he will soon be able to recharge the device.

B. Migration Manager

Based on the information collected by the Context Monitor, the Migration Manager derives energy costs for data communication. Given the data size and the specific communication technology used, the required energy for a transmission can be calculated with the energy model presented in [10]. Based on the composition model, which describes all integral components and their type (UI / none-UI), the Migration Manager identifies which components could be migrated in general. At this time, we assume, that UI components remain unchanged on the mobile device and will not be replaced adaptively.

Analyzing the calculated energy costs for communication and the given device's context information as well as the user requirements, the Migration Manager determines which components have to be migrated to or from the server. The component's state has to be serialized, transferred to the server, de-serialized and the component has to be instantiated with the former state on the server. To lower the transfer overhead between client and server, the component's code itself is not moved to the server. Instead, the server fetches the component code from the repository via the component's ID, known from the composition model. If a component should be migrated from the server to the mobile device, the savings of the communication costs must exceed the transfer costs of the component and its state, if the component has not been instantiated on the client earlier. Thus, the Migration Manager holds information on the components' migration history. Main parts of the Migration Manager shall run on the client to avoid transmitting great quantities of context data to the server for processing.

C. Runtime Environment

The runtime environment is responsible for interpreting the composition model, contacting the component repository, receiving the component code, integrating the components and establishing the needed event channels between the components according to the composition model. Implementations of a CRUISe runtime environment have been developed for server or client side only, recently as a Thin-Server-Runtime [13], running completely within the web browser. However, our approach requires a runtime environment on both the server and the client side, which allows for the dynamic execution of none-UI components on both sides. Thus, a dynamic *Client-Server-Runtime* is currently under development, which provides component integration and instantiation on both the server and the client as well as an event bus (embedding the needed event channels) between server and client to allow for communication among migrated components. Initially, we will utilize a server-side JavaScript executor to run CRUISe components (which typically are JavaScript-based) on the server, supporting further platforms in the future. Access to external services (cf. SA) will be provided in the same manner on server and client, to make the actual location of execution transparent to the components.

V. Conclusion and Future Work

In this paper, we have introduced our proposal for mobile energy-adaptive RIAs briefly. Based on collected context information on the application and the device at runtime, the Migration Manager decides whether to migrate components between server and client to minimize data transfer and, thus, to save energy and lengthen the device's uptime.

We neglect energy optimization on server side within our approach, as we focus on mobile devices. Optimizing energy consumption of component-based applications on servers has been studied within the CoolSoftware project [14], introducing *Energy Auto Tuning* [15].

Our approach has also some limitations. Due to the focus on communication behavior, CPU-intense applications with minor external communication will not benefit much from this proposal. Moreover, it could be difficult to derive a migration strategy for components that consume and publish data of equal quantity, as they will cause high communication costs regardless of where they are executed. Frequent migration of components could result in high overhead costs. This can be addressed by migration policies and initial distribution suggestions for components, derived from experiments run prior to application deployment.

To achieve our aim, we will face the following challenges next: We will survey, whether the Context Monitor could also use prediction technologies (besides runtime monitoring) to determine data traffic or if methods and work from machine learning could be useful as well. Further research is required with regards to migration strategies, clarifying where components should be integrated initially: on the server or on the client, as this impacts the initial data traffic. Finally, we have to complete the implementation of the dynamic Client-Server-Runtime.

To evaluate our approach, we plan a representative user study, which allows us to measure energy consumption for several usage scenarios and communication technologies on current smartphones and to compare our solution with classical mobile RIAs.

Acknowledgment

The work of Johannes Waltsgott is founded by the German Federal Ministry of Education and Research under promotional reference number 13N10782.

References

[1] V. Tietz, S. Pietschmann, G. Blichmann, K. Meißner, A. Casall, and B. Grams, "Towards Task-Based Development of Enterprise Mashups," in *Proc. of the 13th Intl. Conf. on Information Integration and Web-based Applications & Services*. ACM, 2011, pp. 325–328.

[2] Gartner Inc., "Gartner Says Sales of Mobile Devices Grew 5.6 Percent in Third Quarter of 2011," 2012, Mar 22th. [Online]. Available: http://www.gartner.com/it/page.jsp?id=1848514

[3] M. Satyanarayanan, "Mobile Computing: the Next Decade," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 15, no. 2, pp. 2–10, 2011.

[4] E. Cuervo, A. Balasubramanian, D.-K. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making Smartphones Last Longer with Code Offload," in *Proc. of the 8th Intl. Conf. on Mobile Systems, Applications, and Services*. ACM, 2010, pp. 49–62.

[5] CRUISe Consortium, "CRUISe project," 2012, Mar 22th. [Online]. Available: http://www.mmt.inf.tu-dresden.de/cruise/

[6] S. Pietschmann, "A Model-Driven Development Process and Runtime Platform for Adaptive Composite Web Applications," *Intl. Journal On Advances in Internet Technology*, vol. 2, no. 4, pp. 277–288, 2009.

[7] CRUISe Consortium, "CRUISe Mashup Component Description Language MCDL," 2012, Mar 22th. [Online]. Available: http://www.mmt.inf.tu-dresden.de/cruise/mcdl/

[8] S. Pietschmann, V. Tietz, J. Reimann, C. Liebing, M. Pohle, and K. Meißner, "A Metamodel for Context-Aware Component-Based Mashup Applications," in *Proc. of the 12th Intl. Conf. on Information Integration and Web-based Applications & Services*. ACM, 2010, pp. 413–420.

[9] A. Carroll and G. Heiser, "An Analysis of Power Consumption in a Smartphone," in *Proc. of the 2010 USENIX Annual Technical Conf.* USENIX, 2010, pp. 21–34.

[10] N. Balasubramanian, A. Balasubramanian, and A. Venkataramani, "Energy Consumption in Mobile Phones: A Measurement Study and Implications for Network Applications," in *Proc. of the 9th ACM SIGCOMM Conf. on Internet Measurement Conference*. ACM, 2009, pp. 280–293.

[11] D. Fischer, S. Föll, K. Herrmann, and K. Rothermel, "Energy-efficient Workflow Distribution," in *Proc. of the 5th Intl. Conf. on Communication System Software and Middleware*. ACM, 2011, pp. 2:1–2:8.

[12] J. Flinn, "Extending Mobile Computer Battery Life through Energy-Aware Adaptation," Ph.D. dissertation, Carnegie Mellon University, December 2001.

[13] S. Pietschmann, J. Waltsgott, and K. Meißner, "A Thin-Server Runtime Platform for Composite Web Applications," in *Proc. of the 5th Intl. Conf. on Internet and Web Applications and Services*. IEEE, 2010, pp. 390–395.

[14] CoolSoftware Consortium, "CoolSoftware project," 2012, Mar 22th. [Online]. Available: http://www.cool-software.org/

[15] S. Götz, C. Wilke, M. Schmidt, S. Cech, and U. Aßmann, "Towards Energy Auto Tuning," in *Proc. of 1st Annual Intl. Conf. on Green Information Technology*, 2010, pp. 122–129.

A Study on Rating Services Based on Users' Categories

Gianpiero Costantino, Fabio Martinelli, Marinella Petrocchi
IIT–CNR, Via G. Moruzzi 1, 56124 Pisa, Italy
Email: name.surname@iit.cnr.it

***Abstract*—In the Internet age, people are becoming more and more familiar in experiencing online services. In many cases, the customer commits herself and her assets in a business transaction with no (or limited) possibility to test the service/good she is booking/buying. Hence, there is the need to prove the trustworthiness of such services for supporting a user in her choice. Many websites feed the customer with reviews of past users representing their degree of satisfaction. In this paper, we consider a scenario where different services may be grouped together to form packets, and we design and implement a simple procedure through which a customer can choose the packet that best satisfies her expectations. The final choice will be driven both by the qualities of the reviews on the constituting services, and by the customer's personal preference and attitudes. To automatise the procedure, we survey real behaviours of users when they choose a service and give reviews, by obtaining a probabilistic model plugged in our simulator. In particular, we deal with the issue of false review, reported by unfair users that intentionally act malevolently. The simulations results show that our system is robust enough up to a certain number of unfair feedback.**

***Keywords*-Reviewing Systems, Design and Evaluation, Probabilistic Client Model, Unfair Feedback.**

I. Introduction

The availability of a large pool of e-services may lead consumers to face the difficulty of choosing the one(s) that satisfy at best their needs. What generally helps in such situations is a service provider in charge of delivering a list of services, decorated with additional criteria supporting the consumer in her choice. A natural support is represented by rating services, *e.g.* by attaching numerical scores, or textual judgments, summing up the degree of satisfaction of past users towards that service. High scores will encourage the consumer in making her choice, even if the final selection will be influenced also by personal preferences (*e.g.* users will not always choose the hotel with the highest score, as it is probably one of the most expensive).

Here, we consider a scenario in which a broker provides a set of services to different kind of clients. We propose a procedure for rating services through review computation and a simple protocol to offer the composition that best satisfy the client's needs. For our prototype, we rely on a probabilistic client model obtained by reproducing the behaviour of real clients when they give feedback and when they choose services. For designing and implementing such a model, we gather and analyse data from two popular websites. We validate the model through simulations, aimed at testing how the system works in presence of unfair clients that intentionally provide false reviews, a frequent misbehaviour confirmed by recent studies, see, *e.g.* [1].

The paper is organised as follows. Section II recalls related work in the area of rating systems. In Section III, we describe the reference scenario, the procedure for review computation, and the protocol for requesting and experiencing packets of services. Subsection IV-A shows how we derive a probabilistic model both for the client choice and the client feedback. In Subsection IV-B, we present a number of evaluations we have carried out. Finally, Section V concludes the paper.

II. Related Work

The rating of a service (or a product) is kept up-to-date according to algorithms generally built on the principle that the new rating is a function of the old ratings and the most recent review(s) [2]. In simple models, such the one adopted by Ebay prior to May 2008, past and new ratings about the outcome of online transactions between a buyer and a seller contribute in an equal manner to the calculation of the trustworthiness of the seller. More recently, Ebay has started considering only the percentage of positive ratings of the last twelve months. The same temporal window is also used in the Amazon marketplace. Other models combine in a weighted mean the old rating and the newest reviews. Proposals to evaluate such weights are based on, *e.g.* the trustworthiness of the reviewer [3]–[5], the evaluation of the users satisfaction for a set of parameters characterising the object [6], the review freshness, or the distance between the single review and the overall score (as suggested in [2]). Other work, like in [7], [8], suggests to weigh more the reviews given by professionals and less the reviews given by regular users. In our approach, ratings are assigned according to categories of users, as commonly classified in popular websites specialised in services advice. The proposed reviewing system is parametric with respect to the weights to be assigned to past and new feedback. In particular, in this paper, we propose a configuration that is optimal, at least for our scenario, with respect to a percentage of unfair ratings and the speed in achieving reviews values comparable with a set of reference values.

Online reviews posted by users should be considered truthful if supported by a reputation mechanism assessing the trustworthiness of the reviewers. We acknowledge

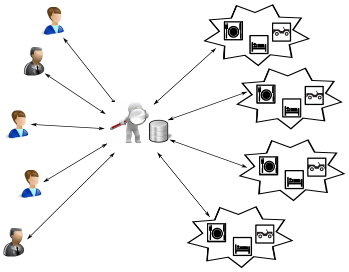


Figure 1. Reference Scenario: Clients-Broker-Packets

research work in the area of immunising reviewing systems against unfair (or incomplete) ratings, *e.g.* [9]–[15]. In particular, work in [16] introduces a new definition of unfairness, by considering two categories of advisers, the first category representing users that intentionally act malevolently, instead the second one representing users that lack of sufficient experience for correctly giving advice. This differentiation allows the authors to propose a two layered filtering algorithm that first detect newcomers with lack of experience, and then classify the remaining advisers according to their credibility. In this paper, rather than proposing a way to cut off unfair ratings, we investigate how robust our reviewing system is, in presence of a certain percentage of unfair ratings.

Work in [17] focuses on feedback selection, and proposes an algorithm to filter past feedback that matches best a user's context. The framework has been tested with real consumers to test its accuracy. Here, instead of dealing with real consumers, we analyse review sets collected from real websites, in order to automatise the behaviours of real users.

Finally, it is worth noticing that recommendation systems have been successfully adopted within large-scale agent-based (social) networks for the selection of trading partners and useful items: as an example, the author of [18] proposes a tag-based recommendation system that maximises some utility function of the users. Also, work in [19] considers how to enable social evaluations and proposes the integration of a cognitive agent and a cognitive reputation model allowing the agent to take decisions in a multi-context environment based on beliefs, desires, intentions, and plans. We acknowledge this area of research as a relevant mean to trade off the subjective attitude of the user's opinion and the community's opinion.

III. Architecture and Scenario

Fig. 1 illustrates our reference scenario, in which an online service broker B provides a list of *packets* P to a client C. Each packet consists of constituting services S^i. As a simple running example, we assume that the client is a traveller willing to book a trip via B. So, C requests accommodation, transportation, and refreshment, and each packet P^j will consist of: hotel S^j_H, car rental S^j_{CR}, and restaurant S^j_R[1]. Hereafter, we let a review range over the set $\{1, \ldots, 5\}$ of real numbers.

The procedure for requesting and experiencing a packet is quite simple:

1) C asks the broker a packet (hotel + restaurant + car).
2) B presents a list of packets, sorted according to the client's preferences. The way in which such preferences are evaluated is explained in Section IV-A1.
3) C chooses the packet whose review best matches her preferences (see Section IV-A1), experiences P, and gives feedback on the services constituting P.
4) B updates the reviews of the single services, and forms a new list of packets for the next client.

We focus on step 4, *i.e.* the computation and updating of the services' reviews. Not surprisingly, we think that the new value should depend both on the more recent reviews and on reviews due to experiences of past users. The following formula generically indicates that the new review is a function f of the old reviews and the last one.

$$R^S_{new} = f(R^S_{last}, R^S_{old})$$

In particular, we propose the next quite simple formula, where R^S_{last} denotes the last review on the service S, R^S_{old} is the old review, and w_{last}, w_{old} are weights ranging over $\{0, \ldots, 1\}$ and $w_{fb} + w_{old} = 1$.

$$R^S_{new} = R^S_{last} * w_{last} + R^S_{old} * w_{old} \qquad (1)$$

The weights are opportunely tuned in order to give more or less importance to history rather than to new feedback.

IV. Validation

In this section, we first characterise in a specific way each actor involved in our scenario. Then, we propose a way to characterise the clients' preferences. In particular, the broker proposes to each client the list of packets in which the first one is the closest to that client's preferences. Also, we present how we derive values R^S of clients' reviews of expression 1. Finally, we propose a number of experimental results, for validating such formula in presence of a percentage of unfair clients that report false reviews.

Our scenario involve a set of clients, a broker, and a set of e-services. In particular:

- The broker is an agent that links services and clients, by following the protocol given in section III.
- The services are hotels, restaurants, and car rentals.
  - In order to validate our proposal, we need reference values for the review of each service in a steady state. For each service, we take as the set of

[1] Here, we simplify the scenario, by considering that all the possible accommodation services (resp., transportation/refreshment services) are represented by a hotel (resp., a car rental/a restaurant).

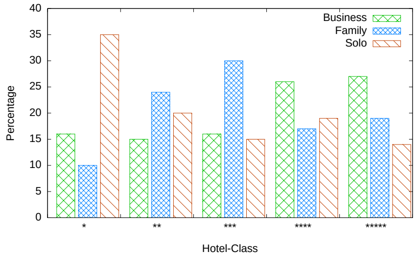


Figure 2. NYC hotels: preference of clients. Percentage of clients choosing NYC hotels, per client typology and hotel class

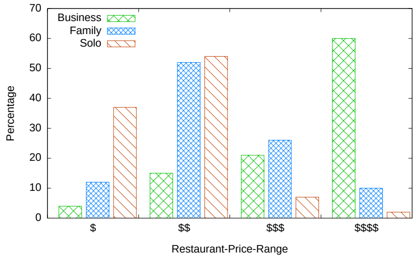


Figure 3. NYC restaurants: preference of clients. Percentage of clients choosing NYC restaurants, per client typology and restaurant price range

review reference values that one surrounding the category reported in the website. As an example, a reference review value for a 5 star hotel ranges over $\{4.51\ldots,5\}$, and for a 4 star hotel over $\{3.51\ldots,4.5\}$.

- Each of the services enters the system with an initial random review value. We justify this choice to test the goodness of our proposal, in terms of proving: 1) if the review values come to results comparable to the reference values (see above); 2) how fast the review computation mechanism is in adjusting the initial random values.

- We consider three categories of clients: solo traveller C_{st}, family C_f, and businessman C_b.

A. Client Model

1) Modeling a client preference: As introduced in Section III, the broker proposes a ranking of packets sorted according to the client's preferences. We assume that three *preference values* are dynamically associated to each client, namely v_h for hotel, v_{rc} for car rentals, and v_r for restaurants. All these three values range over $\{1, \ldots, 5\}$.

We propose to calculate the preference values v_i by considering behaviours of real clients. In particular, we examine popular websites offering travel advices about hotels, restaurants, and car rentals[2].

Regarding hotels, we consider a subset of the 430 hotels in New York City reviewed on Tripadvisor.com. This website allows a user to filter clients' categories, in order to visualize, *e.g.* how many past users of a given category has chosen a particular hotel. Fig. 2 shows the results obtained by our survey. For example, we obtain that, on average, about 27% of the Tripadvisor businessmen users prefer a 5 star hotel, 26% of them choose a 4 star hotel, 16% stay at a 3 star hotel, while 15% and 16% choose, respectively, a 2 star and 1 star hotel.

Regarding restaurants, we consider a subset of the almost 7000 restaurants in New York City revised on Tripadvisor. The website distinguishes them according to the price range, between $ and $$$$. We survey how many businessmen, solo travellers, and families have chosen a restaurant that falls within a particular price range, over a period of time. This leads to the results shows in Figure 3, where it is possible to see that, for example, 60% of businessmen considered in our survey prefer a $$$$ restaurant.

Finally, we consider the website viewpoints.com, giving advice on best car rentals (www.viewpoints.com/Rental-Cars). We notice that the majority of car rentals have a similar number of reviews, meaning that they have been chosen with a similar frequency.

We suggest to assign to each client a preference value v_i in a probabilistic way. For example, a solo traveller will have attached v_h = 1 with probability 35%, = 2 with probability 20%, = 3 with probability 15%, and so on (see figure 2). The same reasoning holds for preference values v_r for restaurants, while, given the results of our survey, we decide to attach to each client v_{rc} = 1 with probability 20%, = 2 with probability 20%, *etc.*. .

Now, we can clarify the way in which the broker sorts the list of packets according to the clients' preferences. Suppose that a businessman asks for a packet (step 1 in the procedure of Section III). First of all, the broker will assign to that businessman v_h^{bus}, v_r^{bus}, and v_{cr}^{bus} in a probabilistic way. Then, B will consider the hotel, the restaurant and the car rental that have obtained reviews closest to v_h^{bus}, v_r^{bus}, and v_{cr}^{bus}. Subsequently, the broker selects the hotel, the restaurant, and the car rental with the second closest values of reviews, and these will form the second packet, *etc.*. . The numerical closeness is in absolute value.

Figure 4 shows an example of a list prepared for a client of category *businessman* whose preference values are v_h^{bus} = 4, v_r^{bus} = 4, and v_{cr}^{bus} = 3. As we can see, the first packet is the one whose components have obtained the review values closest to the client's preference values.

2) Modeling a client review: Once the client has experimented the packet, the broker asks her to provide some feedback. In order to automatise the review computation, we propose a probabilistic feedback model, based on real advices published on Tripadvisor.com. We consider restaurants and hotels in New York City.

On Tripadvisor, each hotel has a set of associated reviews. Reviewers can judge a hotel with five marks: Excellent,

[2] All the surveys refer to data gathered from websites in fall 2011.

1°	H_8= 3.9	R_5= 3.8	C_3= 3
2°	H_6= 3.6	R_3= 3.6	C_1= 3.5
3°	H_2= 4.5	R_6= 4.4	C_8= 4.2
4°	H_1= 4.7	R_4= 4.8	C_3= 1.7
5°	H_5= 2.2	R_7= 2	C_3= 1.5

• • • • •

Figure 4. A list of packets with reviews sorted following the client's preferences. The example shows the list for a businessman with v_h =4, v_r =4, and v_{cr} =3.

Table I

Mark	Feedback Values
Excellent	[4.51 , . . ., 5.0]
Very good	[3.51 , . . ., 4.5]
Average	[2.51 , . . ., 3.5]
Poor	[1.51 , . . ., 2.5]
Terrible	[1.0 , . . ., 1.5]

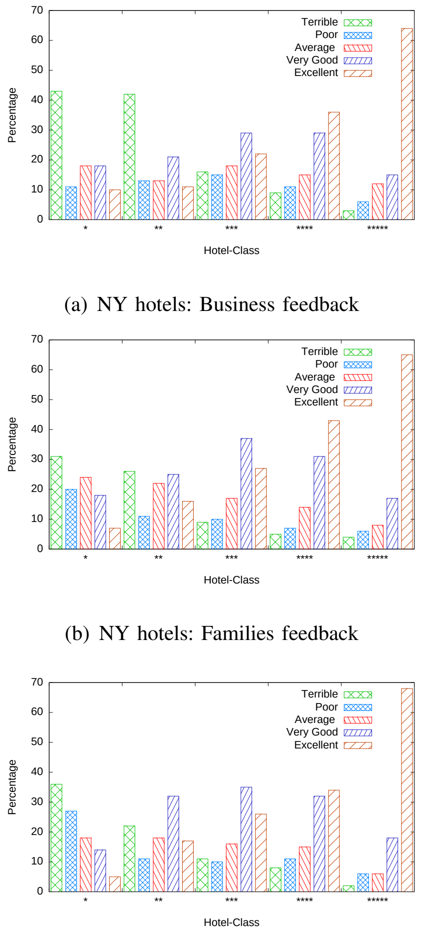
Very good, Average, Poor, Terrible. Reviews may be filtered per client typology, *e.g.* businessmen, families, and solo travellers. Fig. 5 shows the distribution of feedback, per client typology and hotel class. As an example, considering the NYC 5 star hotels, on the totality of 613 businessmen reporting reviews, 393 give an Excellent mark (64%), 92 businessmen a Very good mark (92%), 72 an Average mark (12%), 35 a Poor mark (6%), and 21 a Terrible mark (3%).

Tripadvisor does not allow to filter restaurant reviews according to the client's typology. Thus, we consider a generic traveller. Results of our survey are illustrated in Fig. 6. As an example, we can see that 44% of clients consider a 4$ NY restaurant Excellent, 33% give a Very good mark, 17% think that 4$ NY restaurants are on Average, and 4% and 2% are unsatisfied, giving Poor and Terrible marks.

Finally, reviews on car rentals were not sufficient to derive a feedback distribution. Thus, we decide to consider a uniform distribution of feedback, ranged over {1.0,. . . , 5.0}.

In our system, each service is associated to a default classification (*e.g.* restaurants are classified by price range, and hotels are classified by stars). When a restaurant (respectively, a hotel) is evaluated, a client feedback is probabilistically obtained according to the percentages given in Fig. 6 (respectively, Fig. 5).

For example, a 4$ restaurant is judged *Excellent* with a probability of 44%, *Very good* with a probability of 33%, *Average* 17% and so on. Since we consider as review values real numbers ranged over {1, . . . , 5}, such textual feedback are uniformly mapped to numerical values in intervals as in Table I. These values are the R^S values of expression 1.



(a) NY hotels: Business feedback

(b) NY hotels: Families feedback

(c) NY hotels: Solo travellers feedback

Figure 5. NYC hotels: Clients' feedback. Percentage of clients giving a certain feedback, per client typology and hotel class.

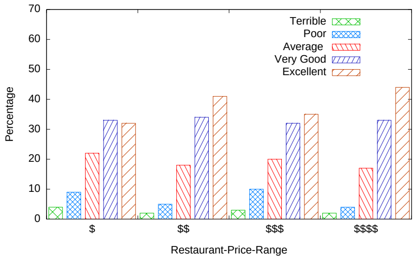


Figure 6. NYC restaurants: Clients' feedback. Percentage of clients giving a certain feedback, per restaurant price range.

3) Unfair clients: Typically, reviewing systems can be altered intentionally by unfair clients. Goal of these users is to post false reviews in order to penalise services. A trivial model is represented by clients who give feedback in a completely random way.

We tackle this issue by considering unfair clients and observing how our system reacts. Here, we adopt a model for the attacker that gives reviews in a probabilistic fashion, and we consider the distribution function got from our

Tripadvisor survey, but in a mirror-like fashion. According to the trend shown in the figures, an *Excellent* mark is given to a high-level service (*e.g.* a 5 star hotel) with high probability. Following the mirror-view strategy, an unfair client gives a *Poor* mark with that same probability.

B. Experimental Results

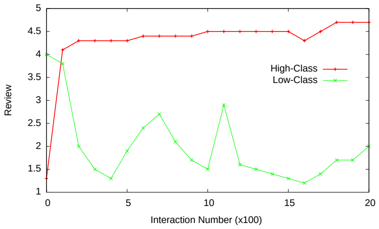


Figure 7. Hotel review ($w_{old} = 0.4$, $w_{last} = 0.6$)

We present some experimental results obtained through a study aiming at characterising the behaviour of our reviewing system. The study is performed implementing an ad-hoc simulator that mimics our framework by letting: 1) the broker propose the list of packets to each client, according to their category and preference values (see section IV-A); 2) the client choose and experience a packet; 3) the feedback be given to each service according to the client's feedback model (see section IV-A2); 4) the broker update the reviews of constituting services according to new and old feedback, following expression 1 of Section III. A number of different interactions is realised in subsequent steps.

The simulator has been developed in JAVA (www.java.com), it is available online[3]. We ran several simulations with different values for w_{old} and w_{last} (see expression 1 in Section III). Tuning the weights, more relevance is given to past feedback R_{old} or to new feedback R_{last}.

1) Fair Clients: Figure 7 shows the review trend in a setting where all clients provide fair feedback. We simulate 2000 interactions: in each of them a client chooses a packet according to her preferences, she experiences and she gives feedback according to her feedback model. Starting by initial random reviews, the services quite quickly obtain reviews very close to the *reference values*. For example, reference values for high class hotels and restaurants are in $\{4.5, \ldots, 5\}$. We can see that the reviews quickly come to comparable values.

2) Unfair Clients: We aim at finding the optimal weights in expression 1 in order to suffer as less as possible from unfair feedback. Thus, we ran several simulations, with different values for weights and different percentages of unfair clients, *i.e.* from 0% to 50%.

In Figure 8 we show the most relevant results we have obtained for a 4 star hotel. On the left column, the review trend is shown in a setting with a low amount of unfair clients (up to the 20% of the totality), while in the right column a higher percentage is considered (up to 50%).

Giving more importance to new feedback, the trend is less stable. Indeed, few new positive (resp., negative) feedback are sufficient for rapidly increasing (resp., decreasing) the service's review values. Hence, an attacker may easily compromise a service, see, *e.g.* Fig. 8(a), and above all, Fig. 8(b), where it is possible to see that a relevant amount of unfair clients can provoke a completely distorted review value. On the other hand, when using very low weights for new feedback (*e.g.* $w_{last} = 0.1, w_{old} = 0.9$, Figures 8(c)-8(d)), the resulting trend is flatter. A flatter trend may affect the disclosure of suspicious behaviours.

The best trade off that we have found between w_{last} and w_{old} is presented in Figures 8(e) and 8(f). A higher importance is given to old feedback. Nevertheless, new interactions are properly considered ($w_{last} = 0.3$ and $w_{old} = 0.7$). Figure 8(f) highlights that these values of w_{last} and w_{old} allow our system to be quite robust even in presence of a high percentage of unfair clients. Indeed, the resulting trend is not affected by substantial modifications.

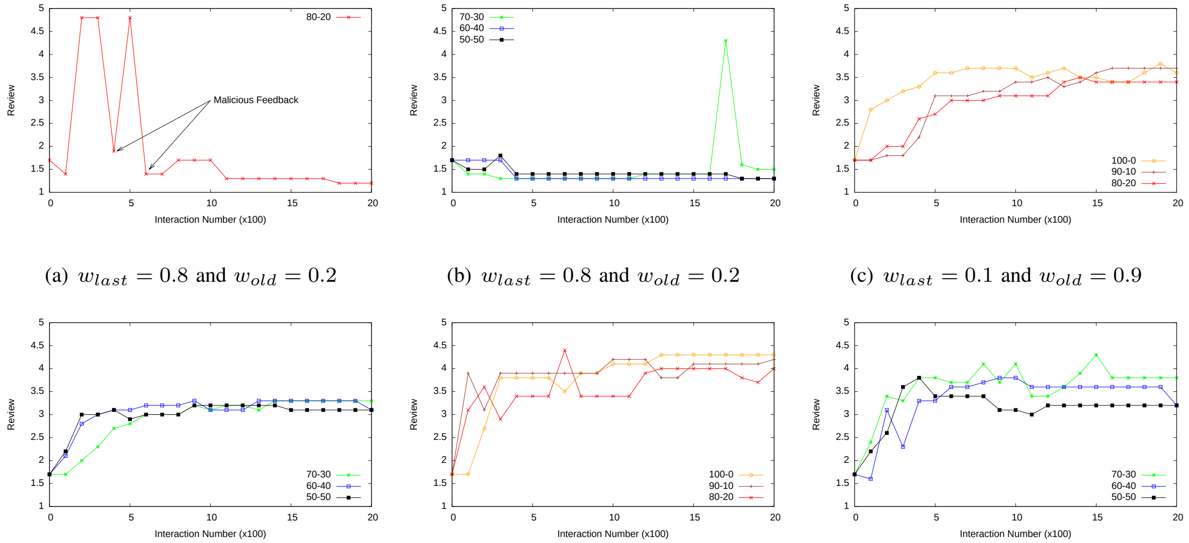
V. Conclusions

We have proposed a rating system for online services. In order to automatise the procedure of review computation, we first collected data from popular websites specialised in clients' reviews. From the analysis of such data, we then derived a probabilistic model of feedback for three kinds of clients: businessmen, families, and solo travellers. The efficacy of the model has been evaluated by simulating a system able to get, as input, feedback of past clients, distributed according to the model that we have derived, and return the updated review value. Simulations show that our mechanism works well up to a certain number of unfair feedback. Also, in our scenario, different kind of services can be composed together and they form packets. Packets are offered to clients according to her preferences, here derived from the analysis of real behaviours of users when they make choice on the Internet.

The surveys that we carried out considers a relatively small number of clients, services, and clients' typologies, but this modeling way could be easily adopted in real world implementations, since many websites specialised in services' reviews usually rely on huge datasets.

We think that other interesting directions could be investigated. First, unfair feedback may lead to a complete distorted review value. Our work could be extended with a proactive component where alarms are raised when something is suspected to go wrong. Secondly, assuming that services initially enter the system with an initial review value fixed in

[3] http://www.iit.cnr.it/staff/gianpiero.costantino/CNR-PersonalPage/Simulator.html



(a) $w_{last} = 0.8$ and $w_{old} = 0.2$

(b) $w_{last} = 0.8$ and $w_{old} = 0.2$

(c) $w_{last} = 0.1$ and $w_{old} = 0.9$

(d) $w_{last} = 0.1$ and $w_{old} = 0.9$

(e) $w_{last} = 0.3$ and $w_{old} = 0.7$

(f) $w_{last} = 0.3$ and $w_{old} = 0.7$

Figure 8. High-class hotel: Review trend varying w_{last}, w_{old}, and the percentage of unfair clients.

accordance with a broker in a business agreement, anomalies between that value and the value calculated with the reviewing system may lead to re-considering the agreement. We leave this for future work based on contracts.

Acknowledgments

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant no 257930 (Aniketos) and under grant no 256980 (NESSoS).

References

[1] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, “Finding deceptive opinion spam by any stretch of the imagination,” in *Human Language Technologies*, 2011, pp. 309–319. [Online]. Available: http://www.aclweb.org/anthology/P11-1032

[2] A. Jøsang, R. Ismail, and C. Boyd, “A survey of trust and reputation systems for online service provision,” *Decis. Support Syst.*, vol. 43, pp. 618–644, 2007.

[3] S. Buchegger and J. Le Boudec, “A robust reputation system for mobile ad-hoc networks,” IC/2003/50 EPFL-IC-LCA, Tech. Rep., 2003.

[4] F. Cornelli, E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, “Choosing reputable servents in a p2p network,” in *World Wide Web*. ACM, 2002, pp. 376–386.

[5] B. Yu and M. P. Singh, “Detecting deception in reputation management,” in *AAMAS*. ACM, 2003, pp. 73–80.

[6] N. Griffiths, “Task delegation using experience-based multi-dimensional trust,” in *AAMAS*. ACM, 2005, pp. 489–496.

[7] T. van Deursen, P. Koster, and M. Petkovic, “Hedaquin: A reputation-based health data quality indicator,” *Electr. Notes Theor. Comput. Sci.*, vol. 197, no. 2, pp. 159–167, 2008.

[8] W. Chen, Q. Zeng, and L. Wenyin, “A User Reputation Model for a User-Interactive Question Answering System,” in *International Conference on Semantics, Knowledge and Grid*. IEEE, 2006, p. 40.

[9] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” in *ACM Conf. on Electronic Commerce*, 2000, pp. 150–157.

[10] A. Whitby, A. Jsang, and J. Indulska, “Filtering out unfair ratings in bayesian reputation systems,” in *Workshop on Trust in Agent Societies*, 2004.

[11] J. Zhang and R. Cohen, “Trusting advice from other buyers in e-marketplaces: the problem of unfair ratings,” in *ICEC*, 2006, pp. 225–234.

[12] Q. Feng, Y. Yang, Y. Sun, and Y. Dai, “Modeling attack behaviors in rating systems,” in *Distributed Computing Systems Workshops*, 2008, pp. 241 –248.

[13] C. Dellarocas and C. A. Wood, “The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias,” *Management Science*, vol. 54, no. 3, pp. 460–476, 2008.

[14] S. Liu, C. Miao, Y.-L. Theng, and A. C. Kot, “A clustering approach to filtering unfair testimonies for reputation systems,” in *Autonomous Agents and Multiagent Systems: Vol. 1*, 2010, pp. 1577–1578.

[15] J. Gorner, J. Zhang, and R. Cohen, “Improving the use of advisor networks for multi-agent trust modelling,” in *Privacy, Security and Trust*, 2011, pp. 71 –78.

[16] Z. Noorian, S. Marsh, and M. Fleming, “Multi-layer cognitive filtering by behavioral modeling,” in *Autonomous Agents and Multiagent Systems - Volume 2*, 2011, pp. 871–878.

[17] W. Zhao *et al.*, “A user-oriented approach to assessing web service trustworthiness,” in *ATC*, ser. LNCS 6407. Springer, 2010, pp. 195–207.

[18] S. Sen, “Finding useful items and links in social and agent networks,” in *Agents and Data Mining Interaction*. Springer, 2010, vol. LNCS 5980, p. 3.

[19] I. Pinyol, J. Sabater-Mir, P. Dellunde, and M. Paolucci, “Reputation-based decisions for logic-based cognitive agents,” *Autonomous Agents and Multi-Agent Systems*, vol. 24, pp. 175–216, 2012.

A Sub-topic Partition Method Based on Event Network

Wei Liu, Dong Wang, Wenjie Xu, Xujie Zhang, Zongtian Liu
School of Computer Engineering and Science, Shanghai University
Shanghai, China
{liuw, ming123, jiex, annie4sh, ztliu}@shu.edu.cn

Abstract—**In TDT (Topic Detection and Tracking) of news stories, a topic usually contains lots of events. Because of small granularity of events, the relationships between events and topic are not closely enough to be distinguished. In this paper, topic and news stories are described by using event networks, and a network clustering algorithm EN-MST based on minimum spanning tree is proposed to discover event communities in the network. Each community is considered to be a sub-topic which could represent an aspect of the large topic as a coarse-grained concept. The experimental results show accuracy and reasonableness by using our method. In our further study, sub-topics obtained by the method proposed in this paper will be adopted to represent news stories in order to distinguish whether a news story belongs to a certain topic.**

Keywords-TDT; topic tracking; event network; community discovery; topic model; EN-MST

I. INTRODUCTION

Topic detection and tracking (TDT) is a research hotspot on information recognition, data mining and organization of news stories, so as to improve the efficiency of useful information acquisition on the Internet. A large topic usually contains many events, and there exist semantical relationships between event pairs, such as casual relations, accompany relations and follow relations. A network of events be formed based on the event relation information to represent the topic. By analyzing the network of events, some events which have more closely relationship are likely to describe one aspect of the topic, so, event clusters in the network are considered to be sub-topics. It is known that VSM (vector space model) is the main text representation method in TDT; however, the main shortcoming of this method is the lackness of semantic information. In this paper, VSM is replaced by event network to represent topics and news stories. A topic hierarchy structure is proposed including topics, sub-topics and events. Events with close relationships are put together by using network community discovery algorithm, to find sub-topics. The sub-topics will be a bridge connecting events and the topics to improve the accuracy of TDT in further study.

The remainder of the paper is organized as follows: Section II introduces the related work on TDT and network community discovery method. Section III discusses the definitions about event and event network and how to construct an event network. Section IV discusses a community discovery method based on MST to obtain sub-topics from an event network. Section V compares experiment results between the objective function in our method and other objective functions. Finally, the conclusions are given in Section VI.

II. RELATED WORK

TDT was proposed in 1996 [1] by the U.S Defense Advanced Research Projects Agency (DARPA). Then, researchers from DARPA, CMU, Dragon system company and Umass [2] began to define the main content of topic detection and tracking, and developed some initial technologies for the solution to these problems. TDT mainly focuses on three tasks[3]: topic tracking, topic detection and new event detection. A topic is considered to be only a collection of news stories. Although hierarchy was proposed for TDT in TDT-2003 [4], the subject of more fine-grained information extraction problems have not yet been considered. In [5], every news story was considered to be an event, and a news probability generation model was proposed. News event generated model including person, places, content and time was automatically built in a unified framework. In [6], a set of sudden outbreak lexical items were extracted in a concentrated time window, and further emergency was identified according to lexical items. In [7], W. Lam extended related words according to the statistical results of vocabulary in news stories, and events were identified by a method similar to Single-Pass cluster. A topic model based on event and event developing relations was proposed by Makkonen [8], but the detail for the method was not provided. Nallapati [9] gave a more specific concept of event identification and event relation extraction with a given topic, in addition, provided the corresponding evaluation methods and test data. Anicic introduced the concept of Event Processing (EP) and a stream reasoning method based on a language called EP-SPARQL [10], which provided syntax and formal semantics to detect compound events.

Another related research hotspot is network community discovery. Girvan-Newman algorithm was proposed by Girvan and Newman [11], whose method was to get communities by finding the edge with the highest score of betweenness and remove it from the network. Newman [12] proposed a weighted network community discovery method based on edge betweenness. He proposed that weighted graphs be mapped into multi-graphs, the betweenness of all the edges be calculated and the largest one be removed in turn until it reaches a reasonable structure. Noack [13] introduced two energy models whose minimum energy layouts represented the cluster structure, one based on repulsion between nodes and the other based on repulsion between edges. Grygorash proposed a graph cluster method based on MST (minimum spanning tree) [14], and removed the edges whose weights were above the threshold to get the community structures.

However, most of the TDT methods were based on traditional VSM which lacks semantic information and the above community discovery algorithms should be improved to be used in our research. In this paper, the content of text is described by using event network instead of the traditional VSM method, and the event network is divided into event clusters by a method based on MST to extract sub-topics from a large topic.

III. Event and Event Network

A. Definitions

In the field of the TDT, a topic is a collection of several related events, including a central event and some other events related to it. A story is a news report closely related to the topic, which contains two or more independent clauses for stating an event. A story is usually a statement of some aspects of a certain topic, while a topic contains all the content of the related stories. An event involves some participants, environment and some other elements. So we introduce the definitions of event, event class and event ontology in [15] which is the foundation of our work.

Definition 1 (Event): event is defined as a thing happens in a certain time and environment, which some actors take part in and show some action features. Event *e* can be defined as a 6-tuple formally:

$$e ::=_{def} < A, O, T, V, P, L >$$

where *A* means an action set happen in an event; it describes the process of the event happens. *O* means objects taking part in the event, *T* means the period of time that the event lasting. The time period includes absolute time and relative time. *V* means environment of the event, such as location of the event; *P* means assertions on the procedure of actions execution in an event; *L* means language expressions. In this paper, we use the event elements *A*, *T* and *V* to represent events. Different events have different elements. A word or a phrase which expresses an event happening can be called the denoter of the event. Each event has an event denoter in text. However, it is not adequate to distinguish events. Such as the *Sichuan earthquake* and *Japanese earthquake*. Although the event denoters *earthquake* are the same, they do not mean the same event due to the different places they occurred in. Therefore, action, time and location are important to represent events.

Definition 2 (Event Class): event class is a set of common characteristics of the event. It can be expressed by *EC*. $EC = (E, C_1, C_2, \cdots, C_6)$;

where *E* is the set of events, which is an extension of the event class. $C_i = \{c_{i1}, c_{i2}, \cdots, c_{im}, \cdots\}$ $(1 \le i \le 6, m \ge 0)$ is the intension of event class, and is the set of common characteristics in the ith element of *E*; c_{im} is one of the common characteristic in the ith element of each event.

In this paper, the definition of event network is proposed as follow.

Definition 3 (Event Network): an event network (*EN*) is a directed acyclic graph that consists of a set of nodes and edges. The nodes are events, and the edges are event relations.

EN::= ($Events$, $Edges$)
$Events = \{e_1, e_2, \ldots, e_n\}$
$Edges = \{<e_i, e_j, r_{ij}>, <e_x, e_y, r_{xy}>, \ldots\}$ $(1 \le i, j, x, y \le n)$
r = {*Correlation*, *Causal*, *Accompany*, *Follow* }

where, *EN* denotes event network, which contains the set of event nodes *Events* and event relations *Edges*. In *Events*, e_i represents an event, r_{ij} represents the relation between e_i and e_j. We define four different relations between events in our event network model, including:

Correlation Relation：if two events appear in the same story and have common event elements, such as time, location or objects, they are correlated.

Causal Relation： if event e_1 causes event e_2, there exists causal relation between e_1 and e_2. Causal is the most important relation between two events. It not only reflects the interaction between events, but also reflects the time sequence of events. For example:

June 1, in the Afghan city of Kandahar, at least 40 people were killed and 60 wounded in an explosion at a mosque.

where *explosion* caused *killed* and *wounded*, so they have the relation of causal.

Accompany Relation：if two or several events almost happen at the same time, they have the relation of accompany. They are often series of actions caused by the same event. For example:

A large truck overturned in the corner, then the electrical tools are knocked out, and clips are thrown out to the ground.

where *overturned*, *knocked out* and *thrown* exist relations of accompany.

Follow Relation: two events have the relation of time sequence, such as *earthquake* and *rescue*, *wake up* and *teeth brushing*.

Definition 4 (Event Ontology). An event ontology is a formal, explicit specification of a shared event model that exists objectively, denoted as EO. The structure of event ontology can be defined as a 3-tuple：

$$EO: \{ECs, R, Rules\}$$

where *ECs* is the set of all events, R indicates all relations between events. *Rules* are expressed in logic languages, which can be used to describe the transformation and inference between events.

In this paper, by using event relations, we connect event instances that appear in text and construct an event network to represent the text. In contrast to word frequency method of VSM, the event network contains more semantic information: events and their relations.

B. Transmission rules of event relationship

The implicit relations will be extracted between events according to the transmission rules in the above event relations.

Causal Relation: The causal relation is transitive. If event B is caused by event A and event C is caused by event B, then, A causes C.

Accompany Relation: If the relationship between event A and event B are accompany, event B and event C are accompany, then there exists accompany relation between event A and event B.

Follow Relation: If event B follows event A and event C follows event B, then C follows A.

In CEC corpus [16] (An emergency corpus which contains 200 Chinese news stories annotated in Semantic Intelligence Lab of Shanghai University.), obvious relations between events have been annotated, and it is difficult to extract all the relationships manually. In order to extract event relations as many as possible to construct an event network, some implicit relations will be annotated by using these transmission rules above.

C. Quantified the relationship between events

While an event network with semantical relationships has been constructed, it is necessary to transform the event network into a weighted network in order to discovery communities of events. The method is introduced in Section IV. Each of the relationship will be mapped into a corresponding weight in the weighted graph, such as the causal relationship, events with causal relations usually describe the developing situation of the topic, which are more important to represent the theme of news stories. Therefore, the weight of causal relation should be larger than other types of relations.

In order to quantify the event relations with weights, The method in article [17] is introduced which was used to calculate the weights between event classes: Choose 200 stories as sample corpus, then add up each pair of event classes on the frequency and calculate the impact factor of them. For one text *d* in the text collection *N*, F_{ei} means the frequency the event class e_i appears in *d*, F_{ej} means the frequency the event class e_j appears in *d*. The formula for calculating impact factor between e_i and e_j is defined as follows:

$$w^d{}_{ij} = \begin{cases} \dfrac{F_{ej}}{F_{ei}}, & if \;\; F_{ei} \neq 0 \\ 0, & if \;\; F_{ei} = 0 \end{cases} \tag{1}$$

If $w^d{}_{ij}$ >1, it is normalized, $w^d{}_{ij}$ =1.

For the whole text collection *N*, the formula for calculating impact factor between e_i and e_j is defined as follow:

$$w_{ij} = \frac{\sum_{d \in M} w^d{}_{ij}}{|\,M\,|} \tag{2}$$

M is the text collection where each text contains event class e_i.

The steps of calculation are: ①Calculate the impact factors of event class pairs in the same text; ②Normalize and calculate the average impact factors of event class pairs in the text collection, which will avoid unreasonableness caused by the large impact factors in a single text.

IV. Community discovery algorithm based on minimum spanning tree

A. Three-layers model structure of topic

In this paper, a three-layers model structure which contains *Event*, *Sub-topic* and *Topic* is proposed. In contrast to a large topic, a news story contains only a few sub-topics (usually less than 4 sub-topics). Therefore, a coming news story can be represented by some event communities and similarity degree between topic and new stories in the sub-topic level, thus to improve the accuracy in TDT.

Topic: a seminal event or activity, along with all directly related events and activities.

Sub-topic: one aspect of the whole topic. A large topic usually contains some sub-topics and each sub-topic focuses on a small aspect of the topic.

Event: a thing happens in a certain time and environment. In news stories, most of the events have implicit elements.

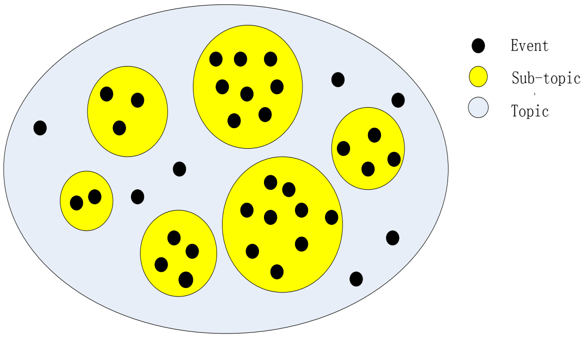


Figure 1. The three-layers model structure of topic.

As Figure 1 shows, a topic contains a lot of events. The granularity of events are too small to describe a topic. Therefore, the sub-topic level which is represented by event communities is proposed to connect events and topic.

B. Communities in network

Community is a description of the close relationships between events. A property of community structure, in which network nodes are joined together in tightly-knit groups and between which there are only looser connections. Currently, there is no recognized evaluation criteria for the community structure. In 2003, Newman proposed the concept of modularity (which is also represented by Q-function) [12]. This quantity is defined as the fraction of edges that fall within communities minus the expected value of the same quantity if edges are assigned at random. Partitioning result depends on the given community memberships and the degrees of vertices. Q-function is defined as follow:

$$Q = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_i k_j}{2m}\right]\delta(c_i, c_j) \tag{3}$$

As to undirected and unweighted graph, if there is connection between node *i* and node *j*, A_{ij} =1, otherwise,

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

A_{ij} =0. $m=\frac{1}{2}\sum_{ij}A_{ij}$ means the edge number in the network. c_i is the community which node *i* belongs to. If $c_i = c_j$, the value of δ function is 1, otherwise, the value of δ function is 0. $\frac{k_ik_j}{2m}$ means the probability of an edge existing between vertices *i* and *j*. Where k_i is the degree of *i*.

Although the Q-function is biased, most of the community partitioning algorithms choose Q-function as a measure means of the clustering quality. A weighted network can be mapped into a multi-graph, and A_{ij} can also represent the weight between *i* and *j*. That means Q-function is also suitable for weighted network. However, Q-function is difficult to obtain the optimal community structure in the sparse graph.

C. Communitiy discovery algorithm based on minimum spanning tree

Most of the community discovery algorithms are for undirected graphs which can not be utilized in event networks. In this paper, an improved algorithm based on minimum spanning tree (MST) is proposed for event clustering.

The traditional clustering based on MST is a splitting algorithm. In the traditional algorithm, the edges with larger weights are removed to form a forest, and every tree in the forest is a cluster. The time complexity of the algorithm is $O(mlogn)$ (*m* is the number of edges, *n* is the number of vertices). Any shape and high dimensional data clustering problem can be processed. However, the MST algorithm has a great complexity, and it is difficult to determine when the algorithm reaches the optimal community structure. Therefore, an algorithm *EN-MST* (*Event Network MST*) using the event relation information of event network for event clustering is proposed. By querying the event ontology, an event network may contain the four types of relations in Definition 3. Because causal relation is the most important relationship, event network will be simplified as follow: remove the relations which are not causal relation and if they have common neighbors, the edges between them can be removed. While utilizing MST algorithm, a simplified network will be faster. An example is shown in Figure 2：

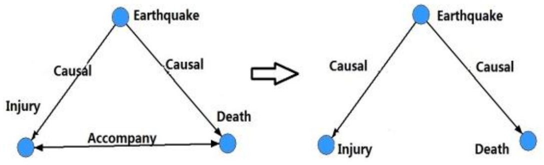


Figure 2. Simplify the event network

In the event network, the greater the weight, the relationship between the two events is more close, and the probability they are in the same community is greater. Therefore, the initial step of the algorithm is to construct a tree with the largest weights which is opposite to MST, that is, the edges with larger weights will be chosen rather than the smaller ones. The adjacent nodes in the tree have close relationship and they may be put in the same community. There exists several branches in a tree and nodes in the same branch are also probably in the same community. However, the granularity of community is too small just by using branch as the criteria, furthermore, the edge weight is not be considered. Perhaps, the community obtained is not accuracy.

In this paper, Q-function method is improved by considering the branches information in the MST, which is defined as follow to avoid the shortcomings of Q-function:

$$Q_{branch} = \frac{1}{2m}\sum_{ij}\left[A_{ij} - \frac{k_ik_j}{2m}\right]\delta(c_i, c_j)\mu(b_i, b_j) \qquad (4)$$

where b_i means the branch that node *i* belongs to. In MST, if there are not nodes whose degrees are more than 3 in the path from *i* to *j*, node *i* and *j* are in the same branch, $\mu(b_i, b_j)$ =1, otherwise, $\mu(b_i, b_j)$ =0.

In the community discovery algorithm EN-MST, there are two main processes. The first one is to generate a minimum spanning tree and the second one is to remove the edges in the tree to get event communities based on the Q_{branch} function.

The steps of EN-MST are described in the TABLE I:

TABLE I. STEPS OF EN-MST ALGORITHM

```
Algorithm: EN-MST
    Step 1. Remove the edges with accompany relation and follow relation;
    Step 2. For the edges with causal relation, query the impact factors between two corresponding event classes from event ontology, and assign them to these edges, thus, get a weighted event network EN';
    Step 3. Calculate all the paths between event pairs. A 2-dimensions matrix B is to save status whether the two events are in the same branch. If node i and j are in the same branch, $B_{ij}$=1, otherwise, $B_{ij}$=0;
    Step 4. Generate a minimum tree ENTree from EN' by the method of Prim.
    Step 5. Remove the edge with the lowest weight, query the matrix B to calculate the value of $Q_{branch}$ and save the previous network condition.
        If (the value of $Q_{branch}$ is higher than before)
            Set $Q_{branch}$ as the largest one;
        Else
            Recover the previous network condition and the next edge is set as the lowest weight ;
    Step 6. Repeat Step 5 until all the edges are checked. Each sub-graph is a community in the event network.
```

Figure 3 shows a minimum spanning tree which is generated from an event network. Three event communities are partitioned by EN-MST is shown in Figure 4.

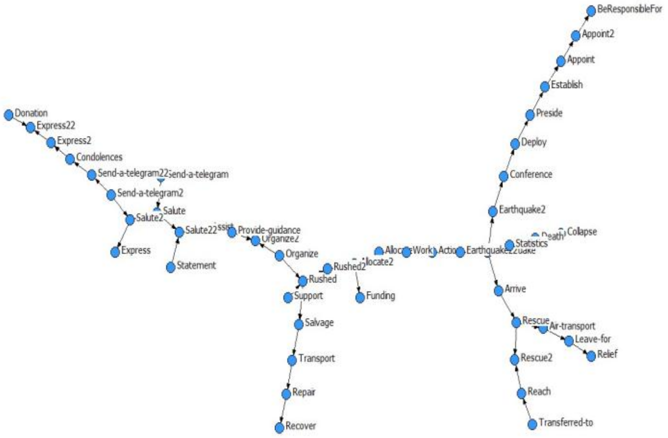


Figure 3. A minimum spanning tree by using prim algorithm

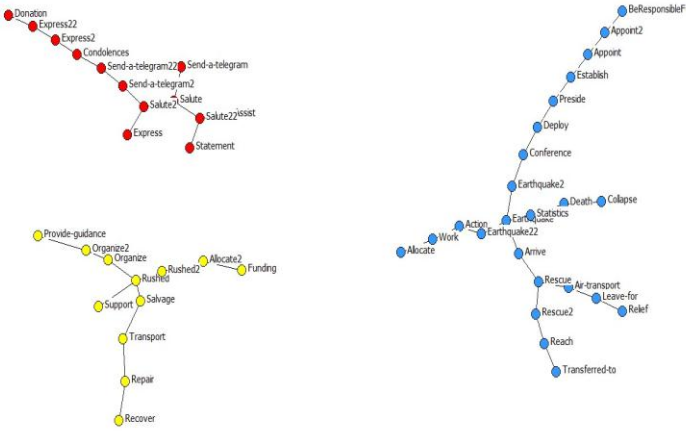


Figure 4. The communities of earthquake event network generated by the EN-MST algorithm

D. *The efficiency of EN-MST*

In event network, semantic information is considered and unimportant edges are removed, which will improve the efficiency of minimum spanning tree algorithm. Compared to Girvan-Newman algorithm whose complexity is $O(m^2n)$, EN-MST only removes the edges according to their weights rather than calculates the edge betweenness, and the number of the edges is the number of nodes minus 1. Therefore, EN-MST have higher efficiency. The time complexity of the algorithm in generating a minimum tree is $O(mlogn)$ (m is the number of edges, n is the number of vertices). In the process of initializing the matrix B, the time complexity is $O(n^2)$, and the time complexity in the removal of edges is $O(m)$.

V. Experiment and Results Analysis

A. *Experimental corpus*

Impact factors between event classes are calculated by CEC corpus including different types of emergencies such as *earthquake*, *traffic accident*, *bromatoxism*, *fire disaster* and *terrorist attack*. Each of the stories from the corpus is considered to be a single topic. For each emergency, four stories are selected to be samples, and sub-topics are partitioned manually according to them. The rest of the stories in the corpus are selected as test corpus.

TABLE II. Sub-topics in samples and their corresponding number of events

Sub-topics	The number of events
Emergency scene	58
The rescue	41
Remedial work	19
Cause of the incident	16
International concern	10
The donation	7

B. *Experimental results and analysis*

In the test corpus, an event network is constructed for each story, and EN-MST algorithm is used for sub-topic partition. In the experiment, the result of the sub-topic partition is represented by a relation between event nodes, that is, two events are either assigned to the same sub-topic or different sub-topics. In a data set with n events, there are $n(n\text{-}1)/2$ event pairs. *RI* is to evaluate the performance of the algorithm according to the correct event pairs which is defined as follow:

$$RI = \frac{\#CD}{n(n-1)/2} \tag{5}$$

where *#CD* is the number of the correct decision on the event pairs. *#CD = A+C*. *A* is the number of the events pairs in which the two events belongs to the same sub-topic both according to the sample and the algorithm result. *C* is the number of the event pairs in which the two events belong to the different sub-topics both according to the sample and the algorithm result. It can be seen that $0<RI<1$，The value of *RI* is larger，the performance of the algorithm is better.

In the test set of the emergency corpus, 20 events are selected at random. The result of the sub-topic partition algorithm is compared to the sample corpus and the *RI* values are calculated by formula (5). In the step of generating a minimum tree, according to MST algorithm, the MST results are not unique, which may cause the variation of community discovery results. Therefore, we repeat this process 10 times, then take the average of *RI* values.

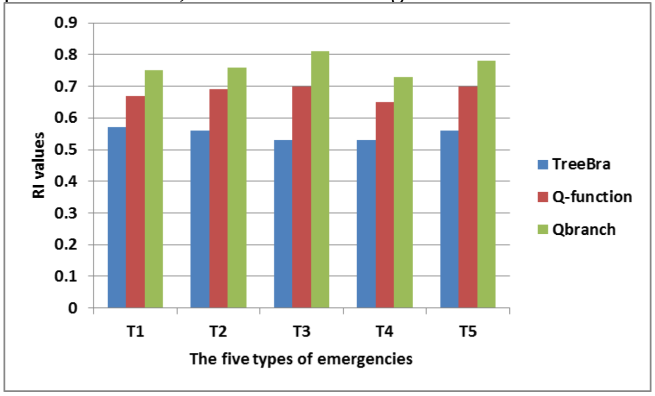


Figure 5. The result of sub-topic parition based on three methods.

TABLE III. THE AVERAGE NUMBER OF SUB-TOPICS USING THE THREE OBJECTIVE FUNCTION

Topics	The average number of sub-topics		
	TreeBra	Q-function	Q_{branch}
Earthquake	5.6	2.3	2.8
Traffic accident	5.4	3.1	3.9
Bromatoxism	4.8	2.5	3.4
Fire disaster	5.8	3.3	3.7
Terrorist attack	6.2	3.6	3.9

In Figure 5, the X-coordinate represents the five kinds of emergency corpus, the Y-coordinate represents *RI* values. The *RI* value by using the method of TreeBra is the lowest. Although it is reasonable to partition community by removing the edges connecting two branches in a MST, the size of each sub-topic is too small,. Sub-topics contain less event nodes compared to the sample sub-topics. Thus, the objective partition result only by choosing the branches in the tree is not corresponding to the sample. Considering the edge weights, the Q-function method has higher *RI* values than TreeBra, because in an event network, the weights of edges play an important part in the removal of node edges and distinguishing whether two events are in the same community. However, the sub-topic partition result by the Q-function is not very reasonable due to ignoring the information of nodes in the same branch. Compared to the above two methods, the algorithm using the objective function Q_{branch} reaches the highest RI value, in which the advantage of TreeBra and Q-function are taken.

Besides *RI* value, TABLE III shows the average number of sub-topics in each event network. A sub-topic may either be the core event which can represent a topic, or the collection of some related events. The granularity of sub-topic is a important factor to evaluating the partitioning result. In the method of TreeBra, the sub-topic number is the most, which means the granularity of the sub-topics is the smallest. In contrast to TreeBra, the size of the sub-topics obtained by the Q-function method is the largest, that means a sub-topic usually contains some events which do not belong to it. In comparison with TreeBra and Q-function, in the method of Q_{branch}, the number of events in each sub-topic is reasonable, and it is the most corresponding to the reality. Therefore, the EN-MST algorithm has the best performance in sub-topic partition.

VI. CONCLUSION AND FUTURE WORK

In this paper, topics and news stories are represented by event networks, which are divided into event clusters by EN-MST to extract sub-topics. A topic hierarchy structure is proposed, which includes topics, sub-topics and events. In comparison of the experiment results, EN-MST gets the highest *RI* values, and the granularity of the sub-topics are the most close to the sample. However, the MST is variation which will influence the results. In our further research, we will find a method to take the place of MST. Similarity calculation using sub-topic information will also be studied to improve the accuracy in TDT.

ACKNOWLEDGMENT

This paper is supported by the Internal Natural Science Foundation of China, No. 60975033 and No.61074135, and Postgraduate Innovation Fund of Shanghai University (SHUCX112161).

REFERENCES

[1] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang, "Topic detection and tracking pilot study: Final report," in Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop, 1998, pp. 194-218.

[2] J. Allan, V. Lavrenko, D. Frey, and V. Khandelwal, "UMass at TDT 2000," TDT 2000 Workshop Notebook,2000, pp. 109-115.

[3] Y. Hong, Y. Zhang, T. Liu, and S. Li, "Topic Detection and Tracking Review," Journal of Chinese Information Processing, vol. 21, pp. 71-87, 2007.

[4] J. Fiscus, "Results of the 2003 Topic Detection and Tracking Evaluation," in Proc. LREC, 2003.

[5] Z. Li, B. Wang, M. Li, and W. Y. Ma, "A probabilistic model for retrospective news event detection," In Proceedings of SIGIR, 2005, pp. 106-113.

[6] G. P. C. Fung, J. X. Yu, P. S. Yu, and H. Lu, "Parameter free bursty events detection in text streams," In VLDB, 2005, pp. 181-192.

[7] W. Lam, H. Meng, K. Wong, and J. Yen, "Using contextual analysis for news event detection," International Journal of Intelligent Systems, vol. 16, pp. 525-546, 2001.

[8] J. Makkonen, "Investigations on event evolution in TDT," In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, 2003, pp. 43-48.

[9] R. Nallapati, A. Feng, F. Peng, and J. Allan, "Event threading within news topics," In Proceedings of the Thirteenth ACM conference on Information and Knowledge Management, 2004, pp. 446-453

[10] D. Anicic, P. Fodor, S. Rudolph, and N. Stojanovic, "Ep-sparql: a unified language for event processing and stream reasoning," in Proceedings of The 20th International Conference on World Wide Web, New York, 2011, pp. 635-644.

[11] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, vol. 99, p. 7821, 2002.

[12] M. E. J. Newman, "Analysis of weighted networks," Physical Review E, vol. 70, p. 056131, 2004.

[13] A. Noack, "Energy models for graph clustering," Journal of Graph Algorithms and Applications, vol. 11, pp. 453-480, 2007.

[14] O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum spanning tree based clustering algorithms," IEEE International Conference on Tools with AI, 2006, pp. 73-81.

[15] W. Liu, W. Xu, J. Fu, Z. Liu, and Z. Zhong, "An extended description logic for event ontology," Advances in Grid and Pervasive Computing, pp. 471-481, 2010.

[16] J. Fu, W. Liu, Z. Liu, and S. Zhu, "A Study of Chinese Event Taggability," IEEE Computer Society 2010, 2010, pp. 400-404.

[17] Z. ZHONG, "Event Ontology and Its Application in Query Expansion," [PhD thesis] Shanghai University, 2010.

Virtual Instrumentation with Mobile Device control for Methane Concentration Measurements

Raul Ionel
Department of Measurements and Optical Electronics
University Politehnica from Timişoara
Timişoara, Romania
e-mail: raul.ionel@etc.upt.ro

Sabin Ionel, Aurel Gontean
Department of Applied Electronics
University Politehnica from Timişoara
Timişoara, Romania
e-mail: sabin.ionel@etc.upt.ro,
aurel.gontean@etc.upt.ro

***Abstract*—This paper presents the design and functionality of a methane (CH_4) concentration measurement system, based on a virtual instrumentation (VI) solution implemented using National Instruments' LabVIEW. It contains a semiconductor type dedicated gas sensor, specific conditioning circuitry and a software program running on a portable computer. An important feature of the proposed implementation is the possibility to transmit calculation results and receive control commands from a mobile phone with internet connectivity. Other features include data logging of concentration parameters and statistical calculations. The instrumentation represents a cost effective solution due to software/hardware adaptability and can be easily extended for monitoring other gases. Experimental results which illustrate the operation of the system in CH_4 contaminated environments are also presented.**

Keywords-LabVIEW; Mobile Phone Control; Gas Sensor; Virtual Instrumentation; Methane Concentration.

I. INTRODUCTION

The use of semiconductor type gas sensors for applications which investigate ambient air pollution levels is a common approach due to several reasons. These sensors combine high sensitivity to target gases, low power consumption, long life, straightforward installation and low cost. Together with an appropriate software application, one can develop a measurement system which can be adapted to monitor one or several gases [1, 2, 3].

The use of combustible gases within industrial and civil buildings represents a permanent explosion and/or fire potential danger. Such disasters happen due to gas accumulations caused by improper device operation or leakage. When concentrations of such gases reach the LEL (Lower Explosion Limit) values, in contact with the air and a fire source (spark, high surface temperature), they can cause explosions which may lead to life and material loss.

Mobile phone applications offer new possibilities of effective remote control and monitoring. They are becoming more reliable, interesting and attractive. Extensive development of such solutions for domains like city car parking, power plant monitoring, SMS (Short Message Service) if monitored parameters are outside allowed limits, SMTP e-mailing or medical condition monitoring, is a current concern. Instrumentation companies offer software and hardware packages for development of mobile solutions for hand-held devices and smart-phones.

Wen et al. [4] describes an interesting tele-monitoring application which records ECG signals, processes the waveforms and sends SMS (Short Message Service) messages to authorized mobile phones if anomalies are automatically detected. The communication between the elements of this system is provided by a Web Server and the TCP/IP protocol.

Another implementation which uses a mobile device for transmission of SMS warnings with the purpose of alerting farmers is proposed by Aziz et al. [17]. Recorded temperature data is processed and with the help of a GSM modem, alerts are sent to mobile phones in case levels exceed accepted limits.

This paper presents a virtual instrumentation which was designed for CH_4 concentration calculation and monitoring. The concept of VI involves a software and hardware ensemble which has the purpose of replacing a stand-alone, dedicated device. The main advantage of this approach is the possibility to exploit the calculus power and performances of the PC on which the software component runs.

Technological innovations together with customer needs for increased functionality in smaller dimensions devices, have caused the spread of VI in domains like: health and medicine, environmental monitoring, remote monitoring, structural investigations, clean energy production, data transmission, industrial control, education, robotics and/or automation.

In related literature, the use of VI is demonstrated by Farhey in [5]. The author presents a solution for monitoring the structure of a bridge. Recorded data are used for evaluating the condition of the construction. The system includes sensors, wireless transmitters and a user-friendly graphical user interface.

Rana and Khan [6] present the complete implementation of a Digital Oscilloscope which uses a NI PCI-6035E card and LabVIEW. This tool is used to study complex signals and includes algorithms for frequency and time domains analysis.

In order to extend the mobility of our application, access to measurement results and program control is available by the use of a web browser on a mobile phone. The user can also choose to store data values and calculations results on

ISBN: 978-1-61208-200-4

the host computer. These options turn the proposed instrumentation into a portable solution which is both reliable, low cost and can be easily adapted to perform different analysis on the measured data [4, 6].

There are two important components which constitute the CH_4 monitoring system.

The hardware includes the TGS2611-C00 methane gas sensor. This sensor satisfies the performance requirements of the UL1484 and EN50194 standards. Specific applications are domestic gas alarms, portable gas detectors or gas leak detectors for gas appliances. It is part of a measurement circuit connected through a data acquisition device to a personal computer. The sensor's analog output is sampled and transmitted to a computer program using USB connectivity. In manufacturer specifications, the use of this sensor is restricted to detecting if the CH_4 concentration exceeds the accepted limit.

The software component uses state-machine architecture for both, control of the measurement circuit and data processing requirements [7, 8]. It was implemented using LabVIEW development environment with the Database Connectivity Toolset and the NI DAQ MX drivers. NI Multisim 11.0 was used for circuit design and analysis. PHP 5.3.5 and MySQL 5.1.54 (embedded in EasyPHP 5.3.5.0) were used for data transmission and application control. Line style flash charts designed by AmCharts were included in the mobile phone user interface. The LabVIEW Web Publishing Tool was used for creating the HTML document which can be accessed using a notebook. In such cases, one has to check if the RunTime Engine software component is installed on the remote computer [10, 11].

Section 2 presents the CH_4 concentration measurement system. This section includes a description of a proposed empirical algorithm for methane concentration calculation. The virtual instrumentation hardware and the software components are discussed. Section 3 presents relevant experimental results. Conclusions regarding both experimental results and future development of the application are mentioned in the final section.

II. METHANE CONCENTRATION MEASUREMENT SYSTEM

An overview of the CH_4 measurement system is presented in Fig. 1. The functionality was tested under laboratory conditions. The measurement circuit contains the CH_4 gas sensor and the proper signal conditioning. This circuit was placed in a laboratory where we could safely feed target gas concentrations to the sensor. Using a data acquisition device (NI USB-6251) as interface between the circuit and the computer running the acquisition software, data is sampled and analyzed. If the user chooses to remotely view the calculation results and control the application, an option for Internet Connectivity is available. In this scenario, raw measured data and calculation results are sent to a remote Web Server with SQL and PHP support. Remote access to all available information requires authentication (using a username and a password) and can be obtained via an html page which resides on the Web Server. For this case we used a Mobile Phone and a Notebook to both view data and control the measurement circuit.

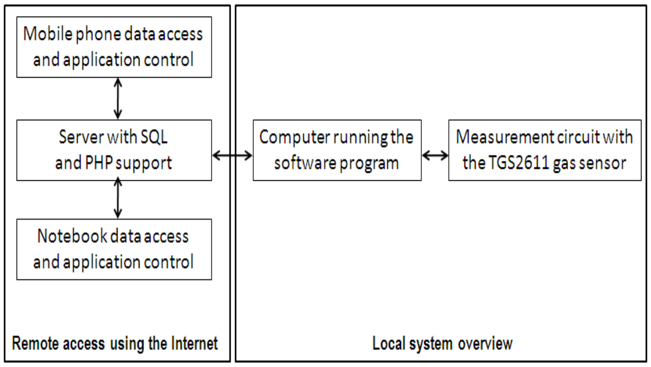

Mobile phone data access and application control
Server with SQL and PHP support
Notebook data access and application control
Remote access using the Internet
Computer running the software program
Measurement circuit with the TGS2611 gas sensor
Local system overview

Figure 1. Overview of the methane concentration measurement system.

The TGS2611-C00 semiconductor type gas sensor is the core of the application. It provides high sensitivity to methane and low power consumption. For this particular application, the sensor was used with the pre-calibrated NGM2611-C13 module. This module includes temperature compensation and meets RoHS regulations. Variations of the electrical parameters for the TGS2611-C00 sensing element are very consistent with concentrations of CH_4 in the surrounding environment [3, 9, 15].

Fig. 2 shows the logarithmic representation for the sensitivity characteristics of the TGS2611-C00 under standard test conditions (manufacturer specifications). One can notice that the sensor internal resistance ratio (R_s/R_o), decreases as target gas concentration increases.

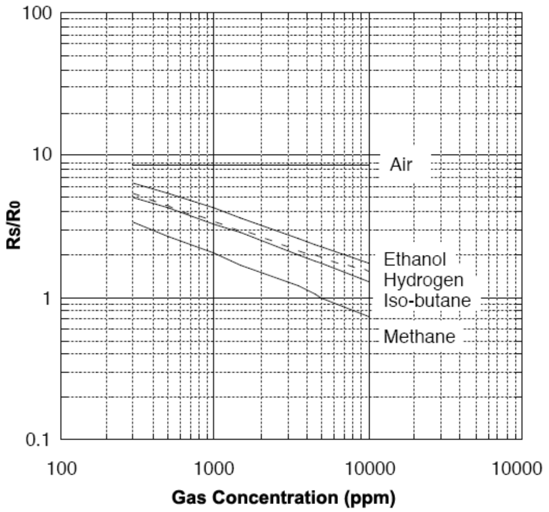


Figure 2. Gas sensor TGS2611 description: sensitivity characteristics.

In the case of methane, the sensitivity characteristics show that R_0 is equal to the sensor resistance R_s when the concentration is 5000 ppm. The recommended concentration values domain is 300 ppm to 10000 ppm. These characteristics are specific for every sensor type. Therefore, one should pay attention to the producer code for each device (#11 in our case) because the load resistor should be chosen in accordance.

The accepted Lower Explosion Limit (LEL) value is 50000 ppm. For this particular case, the NGM2611-C13 module was calibrated to generate an alarm signal when the gas concentration reaches 5000 ppm (or 10%LEL). In practice, this limit is variable due to factors like test conditions tolerances, heat generation inside the sensor enclosure or humidity. If the circuit is operating at recommended parameters, the interval for the accepted alarm limit value is 5%LEL to 20%LEL.

Fig. 3 shows the NGM2611-C13 basic circuit diagram.

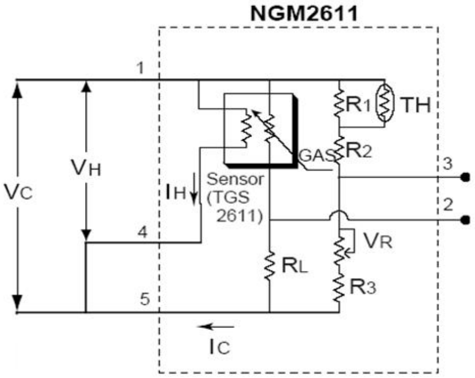


Figure 3. Calibrated module basic diagram.

The operation of the implemented measurement circuit requires a current I_C = 90 mA and a steady V_C = 5 V voltage. As the current flows through the heater, the sensing element is heated and starts to react to the target gas. The voltage divider which contains the sensing element and R_L outputs at pin number 2 an electrical potential which increases with the gas concentration. At the same time, the voltage divider composed of R_1, R_{TH}, R_2, R_3 and R_{VR} (potentiometer resistance) sets on pin 3 the alarm signal threshold which in this case is approximately V_{Alarm} = 2.5 V (alarm threshold). When the voltage on pin 2 exceeds V_{Alarm} the alarm signal is triggered. In this way one can detect if the CH_4 concentration is above the established limit.

A. An empirical relation for concentration calculation

As specified by the manufacturer, the common use of the TGS2611-C00 is for detecting if the CH_4 concentration level exceeds the 5000 ppm limit. This paper extends the application range of the TGS2611-C00 sensor to continuous monitoring of methane concentration. Thus, based on the sensitivity characteristics presented in Fig. 2, an empirical relation between the CH_4 concentration value and the sensor's output voltage was determined. As a first step, the empirical relation (1), between the sensor's internal resistance ratio R_s/R_o and the methane concentration C, was found.

$$\frac{R_s}{R_0} = \frac{D}{C^k},\ for\ 300\,ppm \le C \le 10000\,ppm \qquad (1)$$

The values of the empirical constants are: D = 46.2 and k = 0.45. Fig. 4 shows two graphical representations of the internal resistance ratio (R_s/R_o) versus methane concentration. The logarithmic representation (bottom) is consistent with the appropriate sensitivity characteristic presented in Fig. 2. This means that the constants D and k were accurately determined.

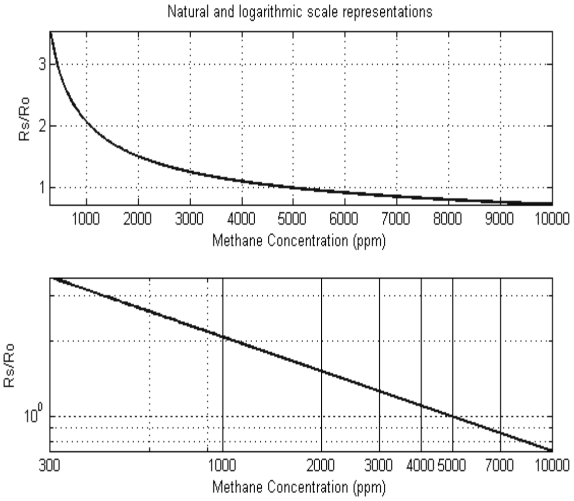


Figure 4. Resistivity ratio versus. methane concentration: natural (upper) and logarithmic scale representation (bottom).

From Fig. 2, the voltage V_{RL} measured on the load resistor R_L can be expressed as:

$$V_{RL} = \frac{V_C}{R_S + R_L} \cdot R_L \qquad (2)$$

For an output voltage $V_{RL} = V_C/2$ = 2.5 V corresponding to a CH_4 concentration C = 5000 ppm, it follows that R_L = R_s at C = 5000 ppm, or R_L = R_0. Thus, (2) can be also written as:

$$\frac{V_C}{V_{RL}} = \frac{R_0 + R_S}{R_0} = 1 + \frac{R_S}{R_0} \qquad (3)$$

Combining (1) and (3), one can determine the empirical relation between the measured output voltage and the methane concentration, C :

$$V_{RL} = \frac{V_C}{1 + 46.2 \cdot C^{-0.45}},\ for\ 300\,ppm \le C \le 10000\,ppm \qquad (4)$$

Finally, from equation (4) one can express the methane concentration as a function of the measured output voltage.

$$C = \left[46.2 \cdot \frac{V_{RL}}{V_C - V_{RL}} \right]^{\frac{1}{0.45}} \qquad (5)$$

Fig. 5 shows the natural scale representation of the CH_4 concentration values as a function of measured sensor output voltage.

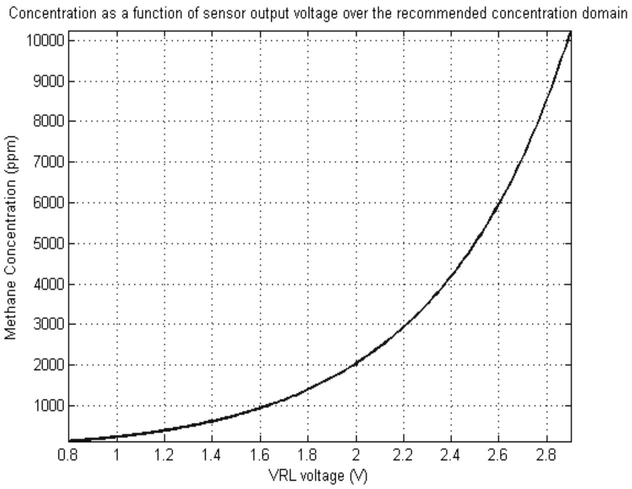


Figure 5. Methane concentration versus sensor output voltage.

B. *Virtual instrumentation hardware*

For this particular application, the data acquisition device must provide 2 analog input lines and the 5 V power line. The multifunction NI USB-6251 device offers the required features. A connector with screw terminals (CB-68LPR) was used to easily access individual signals. Highly selective response to CH_4 can be obtained by eliminating transients by using a filtering material or an appropriate delaying circuit. The latter solution was chosen for this particular implementation. The sensor's response to the target gas stabilizes within minutes (2.5 minutes according to manufacturer specifications), depending on how long it has been inactive. A warm-up alarm prevention RC circuit was implemented with the purpose of delaying the sensor response after power-up. Only when the acquired data is stationary can the correct concentration be calculated [12, 13].

Target gas concentration may fluctuate around the 10%LEL level. A circuit for prevention of disturbing intermittent alarming was implemented. In this way a range for the alarming threshold was created. The alarm is triggered when the sensor voltage exceeds the upper range and lasts until the voltage drops below the lower range.

A final RC circuit was implemented with the purpose of eliminating false alarms caused by the sensor’s reaction to transient interference gases such as alcohol vapors. The recommended timing for the alarm delay is 15 s.

The component values for the TGS2611-C00 sensor were set in order to simulate the case when the R_S value corresponds to a 5000 ppm gas concentration. The values for the NGM2611-C13 module components are set according to specific measurements. The Data Acquisition part was implemented using the MCP601 amplifier as a voltage follower. In this way the sensor voltage is transferred to an input line of the NI USB-6251. The conditioning circuits were designed using LM339 comparators. A single LM339 chip was used in the actual circuit since it contains four separate comparators.

C. *Virtual instrumentation software*

The main features of the proposed instrumentation are:

- The possibility to change the acquisition sampling time from 1 second to 5 seconds. The user can stop the acquisition and resume it without losing the displayed data.
- For ulterior processing, measurement data and calculations results can be saved in a text file on the host computer.
- The user can activate the option for sending data to the server and for allowing remote application control from the mobile phone.
- Calculations are performed and displayed with each measurement. Methane concentration calculated in ppm and percentage, real-time voltage values for the sensor signal and the alarm limit (V_{Alarm}). The concentration values are obtained based on the linearity of the sensitivity characteristics. The voltage values can be studied in order to see if the circuitry is functioning at correct parameters.
- Error messages are displayed if the server or the input lines of the NI USB-6251 cannot be accessed.
- Time domain representations for the sensor output voltage, alarm limit voltage and a running average of the last 4 measured sensor voltage values are presented.

Some practical applications in which these features can be used are gas concentration monitoring in residential buildings, tunnels or underground parking. Also, collected data can be used for statistical calculations which are used for long term studies of concentration evolution.

Fig. 6 presents the basic execution diagram of the software component.

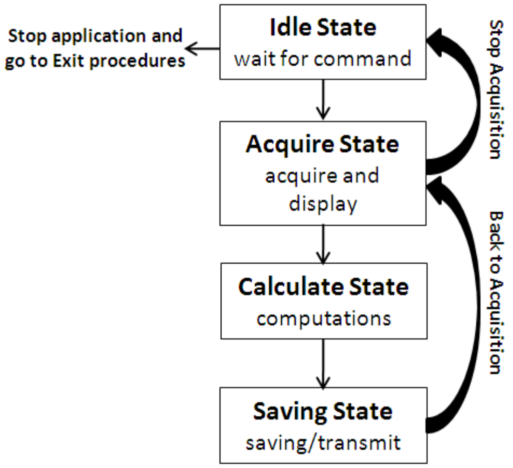


Figure 6. Basic software execution diagram.

The software component was developed using the JKI software add-on state-machine architecture. The advantage of using the JKI state-machine is that starting from a well defined structure, new states for Acquiring Data, Calculations and Data Saving were introduced. We adapted the existing Data Initialize/CleanUp states according to particular needs. Another important issue is the functionality of the front panel buttons when the application is running. Property nodes which disable and enable the front panel buttons can be used as the code execution flows through the states. In this way one can avoid the situation of front panel freezing when the application is in the Acquiring Data state and the user presses the Exit button.

Fig. 7 shows the front panel of the virtual instrument. A short period when the gas concentration exceeds the allowed limit can be noticed on the voltage waveform graph. Operation settings, calculation results and functionality errors indicators are included.

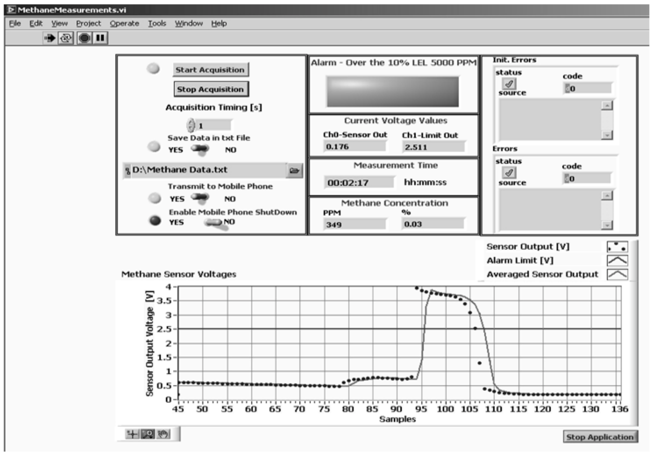

Acquisition Timing [s]: 1
Save Data in txt File: YES / NO
D:\Methane Data.txt
Transmit to Mobile Phone: YES / NO
Enable Mobile Phone ShutDown: YES / NO
Current Voltage Values: Ch0-Sensor Out 0.176; Ch1-Limit Out 2.511
Measurement Time: 00:02:17 hh:mm:ss
Methane Concentration: PPM 349; % 0.03
Init. Errors: status ✓; code 0; source
Errors: status ✓; code 0; source

Figure 7. Front panel of the virtual instrument.

The Database Connectivity Toolset is used to transmit and receive information to/from a server from/to the application running on the software program. Before running the program, an UDL (Universal Data Link) file was created in order to define the communication with the server. This file is used by the DB Tools Open Connection function from the Database Connectivity Toolset. Once the remote connection is successful, the program will be able to transmit and receive data over the Internet. Two tables are used for this particular application. One is used for storage of current measurement data, the other is used for remote commands sent from the mobile phone to the instrumentation [14, 16].

If new measurement data is available on the server, a remote user can connect to the database and view the results. The same server hosts the PHP files which can be accessed using the mobile phone browser. Furthermore, if the remote user works with a portable computer, two connectivity possibilities are available: either by using the PHP files on the server or by using the HTML document created with the LabVIEW Web Publishing Tool on the host computer. The client computer must have the RunTime Engine software component installed. This allows complete control of the main application, if requested by the client computer and granted by the host computer.

When accessing the PHP file on the server, an AmCharts' Flash line graph is loaded. Using the same PHP script, data values (last 12 recorded) and calculation results are read from the database and displayed on the mobile phone screen. If needed, the mobile phone user can stop the main application.

Fig. 8 shows the remote operation of the measurement system using a client notebook (upper image) and the mobile phone (bottom image). In both cases one can notice that the measurement system senses the presence of a higher methane concentration.

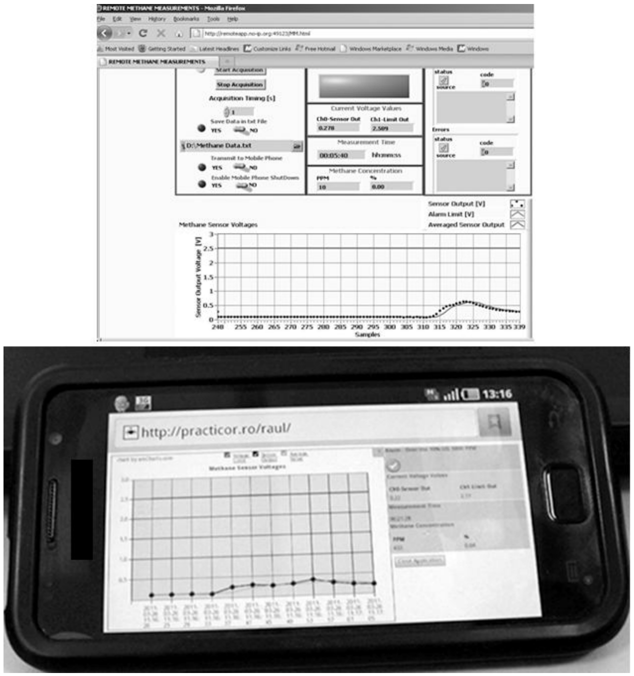

http://remoteapp.no-ip.org:49123/MM.html
Acquisition Timing [s]: 1
D:\Methane Data.txt
Current Voltage Values: Ch0-Sensor Out 0.278; Ch1-Limit Out 2.509
Measurement Time: 00:05:40 hh:mm:ss
Methane Concentration: PPM 10; % 0.00
status ✓; code 0
Errors: status ✓; code 0
http://practicor.ro/raul/
REMOTE METHANE MEASUREMENTS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Most Visited Getting Started Latest Headlines Customize Links Free Hotmail Windows Marketplace Windows Media Windows
REMOTE METHANE MEASUREMENTS
Start Acquisition
Stop Acquisition
Save Data in txt File
Transmit to Mobile Phone
Enable Mobile Phone ShutDown
Sensor Output [V]
Alarm Limit [V]
Averaged Sensor Output
Methane Sensor Voltages
Sensor Output Voltage [V]
3
2.5
2
1.5
1
0.5
0
240 255 260 265 270 275 280 285 290 295 300 305 310 315 320 325 330 335 339
Samples
13:16
Methane Sensor Voltages

Figure 8. Illustrations for the remote operation of the virtual instrument.

III. Experimental Results

Confirmation of the application's functionality was carried out in laboratory conditions. The NGM2611-C13 module was exposed to concentrations which were above and below the accepted LEL value.

Table 1 presents measurement results recorded several minutes after the sensor response has settled, including the short time period when a high CH_4 concentration was recorded. Immediate reaction to the presence of CH_4 inside the sensor's enclosure can be noticed. After the target gas slowly exits, the sensor's response falls to initial values. The values presented in the table were obtained using (5). Since this relation is considered accurate over the 300 ppm to 10000 ppm domain, one can notice that the valid calculations are presented in bold.

TABLE I. Recorded Measurement Values When the Sensor Was Exposed to a Sudden High Methane Concentration

Sensor Output (V)	Alarm Limit (V)	PPM (rounded)	Alarm
1.23	**2.51**	**413**	**0**
1.36	**2.51**	**558**	**0**
1.49	**2.51**	**740**	**0**

1.51	**2.51**	**772**	**0**
1.49	**2.51**	**740**	**0**
1.59	**2.51**	**912**	**0**
1.62	**2.51**	**969**	**0**
1.54	**2.51**	**822**	**0**
1.41	**2.51**	**623**	**0**
1.73	**2.51**	**1207**	**0**
1.84	**2.51**	**1493**	**0**
1.87	**2.51**	**1581**	**0**
1.92	**2.51**	**1737**	**0**
1.96	**2.51**	**1872**	**0**
2.04	**2.51**	**2171**	**0**
2.11	**2.51**	**2467**	**0**
2.17	**2.51**	**2751**	**0**
2.28	**2.51**	**3353**	**0**
2.35	**2.51**	**3799**	**0**
2.41	**2.51**	**4227**	**0**
2.48	**2.51**	**4787**	**0**
2.53	**2.51**	**5232**	**1**
2.58	**2.51**	**5718**	**1**
2.61	**2.51**	**6031**	**1**
2.68	**2.51**	**6833**	**1**
2.72	**2.51**	**7339**	**1**
2.71	**2.51**	**7209**	**1**
2.65	**2.51**	**6477**	**1**
2.61	**2.51**	**6031**	**1**
2.54	**2.51**	**5325**	**1**
2.48	**2.51**	**4787**	**0**
2.41	**2.51**	**4227**	**0**
2.24	**2.51**	**3121**	**0**

Fig. 9 presents the graphical representation of the recorded data presented in Table I. The evolution of the recorded values is presented both as sensor output voltage (using the -*- format) and as CH_4 calculated concentrations (using the -□- format). The voltage limitation of 2.51 V is presented as a dashed line. The Alarm region indicates that the calculated concentration exceeded the 5000 ppm limitation.

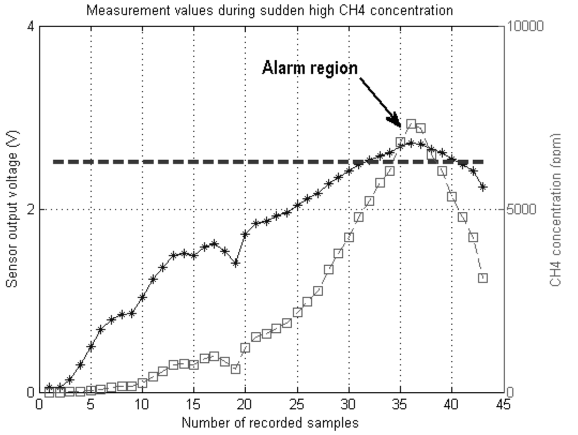


Figure 9. Representation of recorded data and concentration calculations.

IV. Conclusions

In this paper, the design and implementation of a virtual instrumentation solution for monitoring CH_4 concentrations was presented. The application uses the TGS2611-C00 sensor, the NI USB-6251 data acquisition device and the LabVIEW 2009 development environment.

An original and state of the art feature of the proposed system is the possibility to remotely view measurement results and control the operation using a mobile phone with Internet connectivity and Flash script capabilities. Remote access from a client computer, using the LabVIEW Web Publishing Tool, is also possible.

Experimental results showed that the proposed system’s response to the sudden exposure to a high concentration was accurate and fast. Remote monitoring from a mobile phone and a client notebook were successful. This determined the conclusion that the proposed instrumentation has been properly designed and implemented. An empirical formula for concentration calculation was proposed in (5) and was implemented in the software component.

As further development, our goal is to test the precision with which CH_4 concentrations are determined. This can be done by taking measurements in spaces where a predetermined methane concentration is inserted. Furthermore, new options for controlling the instrument using the mobile phone are needed. This will assure complete operation from distance without the imperative need of a computer.

Acknowledgment

This work was partially supported by the strategic grant POSDRU/89/1.5/S/57649, Project ID 57649 (PERFORM-ERA), co-financed by the European Social Fund – Investing in People, within the Sectoral Operational Programme Human Resources Development 2007 – 2013.

References

[1] J. Chou, Hazardous Gas Monitors, McGraw-Hill and SciTech Publishing, 1999.

[2] H. Kohler, J. Roeber, N. Link and I. Bouzid, “New Applications of tin oxide gas sensors I. Molecular identification by cyclic variation of the working temperature and numerical analysis of the signals,” Sensors and Actuators, vol. B 61, Dec. 1999, pp. 163–169, doi: 10.1016/S0925-4005(99)00286-5.

[3] L.D. Dong and D. Sik Lee, “Environmental Gas Sensors,” IEEE Sensors Journal, Vol. 1, No. 3, Oct. 2001, pp. 214–217, doi: 10.1109/JSEN.2001.954834 .

[4] C. Wen, M.F. Yeh, K.C. Chang and R.G. Lee, “Real-time ECG telemonitoring system design with mobile phone platform,” Measurement Journal, Vol. 41, Dec. 2006, pp. 463–470, doi: 10.1016/j.measurement.2006.12.006.

[5] D. N. Farhey, “Integrated virtual instrumentation and wireless monitoring for infrastructure diagnostics,” Structural Health Monitoring Journal, Vol. 5, No. 1, Mar. 2006, pp. 29–43, doi: 10.1177/1475921706057980.

[6] K.P.S. Rana and S.H. Khan, “A DAQ card based mixed signal virtual oscilloscope,” Measurement Journal, Vol. 41, Feb. 2008,pp.1032–1039, doi:10.1016/j.measurement.2008.02.005.

[7] S. Kohout, J. Roos and H. Keller, “Automated operation of a homemade torque magnetometer using LabVIEW,” Measurement Science and Technology Journal, Vol. 16, No. 11, Sept. 2005, pp. 2240–2246, doi: 10.1088/0957-0233/16/11/015.

[8] C.L. Clark, LabVIEW. Digital Signal Processing and Digital Communications, McGraw-Hill, 2005.

[9] E. Llobet, X. Vilanova and X. Correig, “Novel technique to identify hazardous gases/vapors based on transient response measurements of tin oxide gas sensors conductance,” Proc. of SPIE - The International Society for Optical Engineering, Vol. 2504, Jun. 2005, pp. 559–566, doi: 10.1117/12.224147.

[10] G. W. Johnson, LabView Graphical Programming: Practical Applications in Instrumentation and Control, McGraw – Hill School Education Group, 1997.

[11] G. W. Johnson and R. Jennings, LabView Graphical Programming, McGraw – Hill Professional, 2006.

[12] J.S. Bendat and A.G. Piersol, Measurement and Analysis of Random Data, John Wiley & Sons, 1966.

[13] A. Papoulis, Probability, Random Variables and Stochastic Processes, McGraw Hill, 1991.

[14] F.C. Alegria, E. Martinho and F. Almeida, “Measuring soil contamination with the time domain induced polarization method using LabVIEW,” Measurement, Vol. 42, Aug. 2009, pp. 1082–1091, doi: 10.1016/j.measurement.2009.03.015.

[15] L.R. Skubal and M.C. Vogt, “Detection of toxic gases using cermet sensors,” Proc. of SPIE, Vol. 5586, Oct. 2004, pp. 45–53, doi: 10.1117/12.570029.

[16] B. Lin, L. Xiaofeng and H. Xingxi, “Measurement System for Wind Turbines Noises assessment based on LabVIEW,” Measurement Journal, Vol. 44, Feb. 2011, pp. 445-453, doi: 10.1016/j.measurement.2010.11.007.

[17] I.A. Aziz, M. Hasan, M. Ismail, M. Mehat and N. S. Haron, “Remote Monitoring in Agricultural Greenhouse Using Wireless Sensor and Short Message Service”, International Journal of Engineering and Technology, IJET, 9 (2009), pp. 1 –11, doi:10.1109/ITSIM.2008.4631923.

Trust as an Integral Part for Success of Cloud Computing

Felix Meixner, Ricardo Buettner
FOM Hochschule fuer Oekonomie & Management, University of Applied Sciences
Chair of Information Systems, Organizational Behavior and Human Resource Management
Arnulfstrasse 30, 80335 Munich, Germany
f.meixner@ieee.org, ricardo.buettner@fom.de

***Abstract*—Cloud computing has become a hot topic in research in the enterprise and consumer sector. It is clear to everyone that the opportunities and applications of cloud computing are versatile and that cloud computing is an emerging computing paradigm. However when decisions on adopting cloud computing-related solutions are made, trust and security are two of the most critical obstacles for the adoption and growth of cloud computing today. We think there are ways to largely eliminate concerns of potential cloud users by taking advantage of numerous existing technological possibilities, including trust-building measures, like standardization, cryptography, isolation and many more.**

Keywords-cloud computing; security; identity-management; encryption; trust

I. INTRODUCTION

Cloud Computing can be regarded as the most important evolution of the mid 1990's concept of grid computing [1]. In recent years cloud computing clearly became the trend to follow in the IT-industry, providing flexible and scalable software-, platform- and infrastructure-services on demand [2]. However, to fully leverage its potential for cost-savings, cloud computing still has to overcome some major obstacles. As traditional network borders are breaking down at the same time as security threats are increasing, the most important concern about cloud computing are issues of security and trust that have only been partially solved so far.

A lot of literature about cloud computing, trust and security does exist, though most of it is IT-centric [3] [4] [5] [6]. What is less examined and documented is the human perspective that examines the shortcomings of cloud computing, people's expectations and anxieties as well as psychological aspects. This paper's objective is to focus on both perspectives, IT and human and try to narrow the gap between both by offering a state of the art overview of mechanisms that help secure the use of cloud computing and thereby create trust in cloud computing. The research question is: Can cloud computing gain enough trust from its users and customers to be even more successful and become an indispensable utility like the power grid?

Our approach to this subject included research on the history and state of cloud computing today, thereby identifying trust and security as the most critical factors of success for future growth and adoption. With these findings in mind, our research was refined on trust and security in cloud computing and its supporting and control mechanisms. The research methodology included investigating multiple of the most relevant online scientific journals databases (Springer Link, JSTOR, ScienceDirect, Elsevier, IEEE Xplore Digital Library and ACM Digital Library).

The remainder of the paper is organized as follows: In Section II we recognize related work. Then the paper gives an insight into the history, different types and sources of trust in non-technological fields and ways in Section III. These fields include trust in general, in psychological and in economical aspects. The paper outlines the difference between party trust and control trust and sets up a framework for trust that is transferred to Section IV, where the framework is mapped to cloud computing technology. The paper continues with Section V by describing various types of technology aiming to enhance user's and decision makers trust in cloud computing. Finally, in Section VI, we draw the conclusion and provide recommendations for future work and show the need for optimizing existing trust infrastructure and mechanisms.

II. RELATED WORK

In his article "Cloud Computing", Brian Hayes discusses the trend of moving software applications into the cloud and the related trust privacy, security, and reliability challenges [7]. E. Pearson focuses on privacy challenges as important issues for cloud computing, both in terms of legal compliance and user trust and says that it needs to be considered at every phase of design. He suggests key design principles for software engineers and argues that privacy must be considered when designing any aspects of cloud services, for both legal compliance and user acceptance [8]. The article "A View of Cloud Computing" defines classes of utility and cloud computing and creates a ranked list of critical obstacles to adoption and growth of cloud computing. The list includes availability, data lock-in, data confidentiality and auditability as the top three factors for adoption [9]. M. Mowbray and S. Pearson of HP Labs in their paper "A Client-Based Privacy

Manager for Cloud Computing" state that processing sensitive user data in the cloud poses a significant barrier to the adoption of cloud services and that users fear data leakage and loss of privacy. Mowbray and Pearson describe a client-based privacy manager that helps reduce this risk as well as providing additional privacy-related benefits by reducing the amount of sensitive information sent to the cloud [10].

III. Concepts, Types and Sources of Trust

People have been aware of the concept of trust for quite a long time. In fact, it is as old as the history of man and the existence of human social interactions [11]. The majority of literature and studies about trust comes from classic disciplines like philosophy, psychology and economics, all of which concentrate on exploring a general understanding of trust. This paper focuses on trust in cloud computing, by referring to these studies that explain classic forms of trust alias offline trust.

Philosophy traces the concept of trust back to the ancient Greek. They believed that people trusted others, only if they were confident that the others feared detection and punishment enough to deter them from harming or stealing.

Psychology focuses on interpersonal trust and agrees that it was an especially important concept in psychology and vital to personality development (Erikson, 1963) [12], cooperation institution (Deutsch, 1962) [13] and social life (Rotter, 1980) [14]. Rotter gave a frequently cited definition of interpersonal trust as "an expectancy held by individuals or groups that the word, promise, verbal, or written statement for another can be relied on [14]." He has also proven through experiments, that trust has positive consequences to people and society overall.

Economics study trust intensively in organizational contexts. Among other factors it is considered a predictor of satisfaction in organizational decision-making. It was also recognized that trust is able to reduce the cost of both intra- and inter-organizational transactions and able to enhance business performance [15]. Trust, defined as "a willingness to rely on an exchange partner in whom one has confidence", assumed an essential role in establishing and maintaining a long-term relationship between sellers and customers [16].

It can be stated already, that trust is a complex, subjective and abstract concept that is difficult to define. You can find many definitions of trust in literature substituting it with credibility, reliability or confidence. The Oxford English Dictionary in 1971 defines trust as "confidence in or reliance on some quality or attribute of a person or thing, or the truth of a statement''. Mainly though it is a mechanism reducing social complexity on the one hand, but causing vulnerability towards something or somebody on the other hand.

In an article regarding e-commerce, Tan and Thoen considered party trust, control trust and the duality between trust and control as important concepts [17]. Party Trust means trust in the other party. It is subjective and has both an action and an information perspective. Mayer et al. define it as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other party will perform a particular action important to the truster, irrespective of the ability to monitor or control that other party [18]." Control Trust means the trust that is created by a control mechanism. It tends to be more objective than party trust. If there is not enough party trust in a situation, an instance of control trust should be used to increase the overall level of trust. For example, getting a receipt at the dry cleaners stating how many pieces of clothes you handed in, increases your level of trust to get all the pieces back later on.

Psychology was found to one of the most important aspects of trust, which is why it is helpful to have a framework of criteria on how trust is generally observed. Using this framework it will then be possible to draw comparisons between offline trust, in the before described sense, and online trust in the field of technology and cloud computing. According to the overview of Wang and Emurian [11] most researchers study four characteristics of trust:

1. *Trustor and trustee*

A trusting relationship always consists of a trusting party (trustor) and a party to be trusted (trustee). "The development of trust is based on the ability of the trustee to act in the best interest of the trustor and the degree of trust that the trustor places on the trustee"[11].

2. *Vulnerability*

The concept of trust only works and is needed in environments where vulnerability, uncertainty and risk are involved. A trustor relies on the trustee not to exploit his vulnerabilities.

3. *Produced actions*

"Trust leads to actions, mostly risk-taking behaviors. The form of the action depends on the situation, and the action may concern something either tangible or intangible [11]."

4. *Subjective matter*

In every case trust is a subjective matter. Each individual regards trust differently on a case-by-case basis being influenced by personal and situational factors.

IV. Trust in Cloud computing Technology

As the introduction of the paper says, some of the major concerns in cloud computing are trust and security. Trust is one of the most critical obstacles for the adoption and growth of cloud computing. Therefore, in this section we will not only refer to the framework with the four characteristics of trust we have just laid out in the preceding chapter, but go beyond this and include security as an object of study, which interacts bilateral with trust.

1. *Trustor and trustee*

The cloud also relies heavily on the concept of trustor and trustee parties to establish trusting relationships. The difference is that with online trust, the distribution of roles is narrowed down to the cloud service provider being the trustee and the cloud service customer or end user being the trustor.

2. *Vulnerability*

The count of vulnerabilities enterprises face in cloud computing are innumerable. In the digital age of software bugs and ideological hacking groups such as “anonymous” and “LulzSec”, the news are full of exploited vulnerabilities in the Internet. They reach from inadvertent loss of privacy and data theft, to loss of reputation and therefore money. Together, these reasons contribute to the necessity of trust in an insecure and hostile online world.

3. *Produced actions*

Customer’s trust in cloud service providers can generate a couple of desired actions. An enterprise starts using a cloud service and shares its private and precious data with the cloud computing provider. On top of that, an enterprise might be confident to even pay for the cloud service and continue using it on a regular basis.

4. *Subjective matter*

Trust in cloud computing and technology is fundamentally as subjective as its offline counterpart. Again each individual and enterprise has different affections and preferences regarding technology that influences the level of trust towards cloud computing.

Meanwhile even more frameworks regarding trust in cloud computing exist. For example, a recent study from the University of Adelaide showed how to determine the credibility of trust feedbacks. In their paper “Trust as a Service: A Framework for Trust Management in Cloud Environments” they implement the Trust as a Service (TaaS) framework to improve ways on trust management in cloud environments [19].

V. CREATING SYSTEMIC TRUST THROUGH IT TECHNOLOGY

In a world wide web and in clouds of anonymity personal trust is a trait that is very hard to find. Therefore, cloud computing has to earn the trust of enterprises, decision makers and users, by relying on other forms of trust. Fortunately, there are many methods to create systemic trust by means of control mechanisms and help of modern virtualization and security technology.

The next sections follow and expand a proposal for a reference deployment model to eliminate user concerns on cloud security by Zhao, Rong, Jaatun and Sandnes [20]. The model deals with security related issues in cloud computing and proposes five service deployment models to address these issues. The proposed model provides different security related features to address different requirements and scenarios. While some scenarios of the deployment model have multiple valid solutions at hand, others have not yet been entirely solved. Keeping the model in mind it is used as a basis and expanded with some similar, but more practical solutions towards a trusted and secure enterprise cloud:

A. Separation, Isolation and Multi-Tenancy

B. Availability and Reliability

C. Data and Service Migration

D. Cryptography

E. Contractually Fixed Agreements

F. Certifications, Standards Compliance and IT Service Quality

G. Transparency

A. *Separation, Isolation and Multi Tenancy*

Some central mechanisms of increasing importance are identity management and access control. They fit into the category of separation, isolation and multi-tenancy. In contrast to applications and services hosted in-house, proper access management is a must-have. As soon as enterprises decide to use more than one cloud computing service, the challenge rises quickly, due to a couple of issues. Users have to deal with an inflation of credentials, thus increasing the risk of simple and re-used passwords for multiple services. The responsible IT-Managers cannot oversee the access rights of employees or users that are spread across multiple cloud service providers. This fact leads to difficulties in access control management, especially if changes in responsibilities or personnel take place, or an employee resigns. This decentralized identity management also makes central logging of access much more difficult.

A solution to this issue could be to recentralize identity management and access control back into the enterprise by means of single-credential and single-sign-on solutions. A single-credential solution uses a master identity store, either replicated to the cloud, or queried by the cloud service provider, for example via Lightweight Directory Access Protocol (LDAP). A Single-Sign-On solution leverages the single-credential solution and requests authentication from the user only once at the first login. Subsequent authentications to cloud services are automated via asymmetric encryption mechanisms such as Public Key Infrastructure (PKI) using the trust model of certificate authorities (CA). These underlying mechanisms are transparent to the user. Both solutions require an effective protection of the central identity store, as a theft of those credentials provides potential access to all cloud services, granting access based on single-credential or SSO solutions [21].

In their article “Isolation in Cloud Computing and Privacy-Enhancing Technologies” N. Sonehara, I. Echizen and S. Wohlgemuth discuss the common issues around data leakage and loss of privacy [22]. They see isolation as a special kind of privacy protection mechanism, which avoids information exchange between cloud services through their users. Furthermore, isolation should be able to hide the objectives of cloud-users from the cloud service provider. They agree with Ambrust et al. 2010 [9] that the most current and common security mechanism in today’s clouds, to reach the goal of isolation, is primarily virtualization. Ambrust states “It is a powerful defense, and protects against most attempts by users to attack one another or the underlying cloud infrastructure. However, not all resources are virtualized and not all virtualization environments are bug-free. … Incorrect network virtualization may allow user code access to sensitive portions of the provider’s infrastructure, or to the resources of other users. These challenges, though, are similar to those involved in managing large non-cloud data centers, where different applications need to be protected from one another. Any large Internet service will need to ensure that a single security hole doesn't compromise everything else [9]." Due to such flaws in technology, it is important not only to rely on a single mechanism to provide trust and security, but to interlink and connect with other mechanisms, as explained in the following sections.

B. *Availability and Reliability*

Some of cloud computing's key requirements for information security are availability and reliability. Data centers and cloud services should be designed for scalability and performance as well, and limit the necessity of human interaction [23]. Nonetheless we have seen a number of complete datacenters outages in the recent past, including market leaders such as Amazon and Google. Undheim, Chilwan and Heegaard focus on four different types of failures, namely failures in the power distribution or cooling, network failures, management software failures and server failures [24]. For all types of potential failures there are mechanisms in place that help to reduce the availability- and reliability risks to a minimum level. Two of the four mentioned types of failures were picked, and related work was investigated:

Regarding network failures, Gill, Jain and Nagappan present a large-scale analysis of failures in a data center network [25]. Their key observations state that data center networks are already reliable, especially because of their highly redundant design. Nevertheless, there is room for improvement in some areas. They state that load balancer reliability and the effectiveness of network redundancy have to be improved to mask the impact of network failures from applications. Further they recommend separating the network control plane from the data plane to avoid undesirable interference between application and control traffic.

Venkatesh and Nagappan study server failures, hardware repairs and reliability for large cloud computing datacenters and present a detailed analysis of failure characteristics, as well as a preliminary analysis on failure predictors. They state that "8% of all servers can expect to see at least 1 hardware incident in a given year and that this number is higher for machines with lots of hard disks. … Chances of seeing another failure on the same server is high. We find that the distribution of successive failure on a machine fits an inverse curve. … We also find that the location of the datacenter and the manufacturer are the strongest indicators of failures, as opposed to age, configuration etc. [26]." In ongoing work they are working on models for server reliability, including replacing hard disk drives (HDD) with solid state drives (SDD) for better reliability.

Now that we have given an insight into various types of failures, we want to show a conceptual and simple solution design, to circumvent all types of failures that jeopardize availability and reliability of cloud services. The reference deployment model of Zhao, Rong, Jaatun and Sandnes [20] corresponds with the central point on Ambrust's [9] top ten list of obstacles for growth of cloud computing, namely availability + business continuity. Their solution is to use multiple cloud service providers, as they describe in their reference deployment model. The model builds an availability model on top of at the best already redundantly designed cloud infrastructure, adding an extra layer of redundancy of its own. The model achieves this by meeting the following three requirements:

- Get two independent cloud service providers offering equivalent data processing services and two independent cloud service providers offering equivalent data storage services.
- Data replication between both data storage providers is bidirectional and transparent to the user.
- Both data processing services must have access to both data storage services, assumed authorization is granted.

"The Availability Model imposes redundancy on both data processing and cloud storage, hence there is no single point of failure with respect to data access. When a data processing service, or a cloud storage service experiences failure, there is always a backup service present to ensure the availability of the data [20]."

All of the above clearly shows that availability and reliability can be established in multiple and redundant ways, and, therefore are able to contribute to establishing trust in cloud services.

C. *Data and Service Migration*

Another concern of cloud users is potential lack of long-term service viability and, as a result, the inability to get the data, once placed there, out of the cloud, due to data lock-in with one cloud service provider. In this scenario users would be forced to stay with their cloud service provider, who might request premium prices and thus discourage potential customers to use the cloud service at all. They would only use it, if they really had to, or if they were assured that their data could freely be migrated to other cloud service providers.

Hao, Yen and Thuraisingham consider the problem of service selection and migration in a cloud and developed a framework that simplifies service migration. It also includes a cost model and a genetic decision algorithm to discuss tradeoffs of that matter and find the optimal service migration decisions. In their opinion the important issues surrounding the paper are: "It is necessary to consider the infrastructure support in the cloud to achieve service migration. The computation resources (computer platforms) in the cloud need to be able to support execution of dynamically migrated services. We develop a virtual machine environment and corresponding infrastructure to provide such support. … It is also essential to have a strong decision support to help determine whether to migrate some services and where to place them. The consideration involves the service migration cost, consistency maintenance cost, and the communication cost gains due to migration. We develop a cost model to correctly capture these costs and help determine the tradeoffs in service selection and migration in clouds. Then, we use a genetic algorithm to search the decision space and make service selection and migration decisions based on the cost tradeoffs... [27]."

With their reference deployment model Zhao, Rong, Jaatun and Sandnes go a bit further by stating: "a model that can ensure the capability of migrating data from one cloud to another is imperative… [20]." They demonstrate an abstract model where "the migration of data is guaranteed". The model utilizes a data processing service through which users process their data and that is capable of migrating data from one cloud storage service to another. The model achieves this by meeting the following three requirements:

- There is a Cloud Data Migration Service that can interact with the Cloud Storage Service that keeps users' data for exporting users' data.
- There is a second Cloud Storage Service that allows users to import and export data.
- Two independent cloud providers should provide the two Cloud Storage Services.

Hirofuchi, Ogawa, Nakada, Itoh and Sekiguchi are fulfilling this migration model and believe "the next stage for IaaS cloud technology is cloud federation … users can easily deploy their applications on any IaaS cloud providers in the same manner, and transparently relocate them to other providers on demand [28]." They back up their proposal with an "advanced storage access mechanism that strongly supports live VM migration over WAN. It rapidly relocates VM disks between source and destination sites with the minimum impact on I/O performance. It is implemented as a transparent proxy server for a storage I/O protocol … which can be integrated into SAN services in datacenters. This means that the proposed mechanism is independent of VMM implementations [28]." This counters the risk of data lock-in with a particular provider, while still enabling users to select the most appropriate provider any time with the framework of Hao, Yen and Thuraisingham.

The solutions and proposals in [20][27][28] correspond to the second central point on Ambrust's [9] top ten list of obstacles for growth of cloud computing, namely data lock-in. He thinks standardization of APIs and compatible software enable a surge or hybrid cloud computing. Offering different cloud service selection and migration models, as well as standards, can be used to increase trust in cloud computing.

D. *Cryptography*

One common way to preserve key requirements, such as confidentiality and integrity in computing, is to encrypt data before, during and even after transport through the Internet for secure storage. As the cloud service provider has access to the data of all its customers, and may offer it, inadvertently or deliberately, to third parties, there is an urgent need for data encryption. One way to conduct this measure is by using combinations of encryption mechanisms. The trust-building and underlying technique used is pre-egression or pre-internet encryption (PIE). This simply means, encrypting data with your own encryption keys before sending it to the cloud. The encryption keys are in possession of the data owner only and are unknown by the cloud service provider or any 3rd party. After the data is encrypted locally it will leave the local premises and transit through the Wide Area Network (WAN). The cloud service provider should not only offer a tunneled and encrypted transit through the network to the storage destination in the cloud. He should also offer encrypted storage of the data. However, since the cloud service provider knows the encryption keys to those tunnels and storage, the only secure method of processing data is the aforementioned PIE.

Pushing the idea of end-to-end encrypted data even further, is the concept of homomorphic encryption. It can be used to conduct mathematic operations on encrypted data without decrypting it [29]. The major and still unsolved downside to this approach is the immense computing power needed to process the encrypted data and limited support for computing operations, which is why this concept is almost unheard of in the public discussion about cloud trust and cloud security.

E. *Contractually Fixed Agreements*

As stated earlier in the text, trust can be established by establishing control mechanisms. One example of those control mechanisms is Security Service Level Agreements (SSLA) sometimes also referred to as Protection Level Agreements (PLA). They include contractually fixed security restrictions, compliance checks, as well as security information and event management (SIEM). They can be compared to general terms and conditions a company bases its contracts on or to an acceptable use policy (AUP) and are the only legal obligation of the cloud service provider. However, as of today, besides the technical standardization, there are no publicly defined standards yet in the field of information rights management, secure virtual runtime environments and externalization of identities [30][31].

F. *Certifications, Standards Compliance and IT Service Quality*

Online trust needs a solid and justified foundation to build upon. There are a number of trust-building measures in the field of standards compliance and certifications, three of which we find particularly appealing.

The first trust-building measure that should help choose the right cloud service provider is certifications. Looking at geographical boundaries, there is the Cloud Security Alliance (CSA) in the US and the Federal Agency for Information Security (BSI) in Germany. Both support an initiative called EuroCloud Star Audit that provides a seal of quality for Software-as-a-Service (SaaS), one of the three subdomains of cloud computing. It focuses on topics like data security, data privacy, drafting of contracts and compliance on the one hand, on the other hand, topics such as professional IT management, transparent and comprehensible processes, encryption, backup, archiving, exit-strategy, service level agreements, performance and many more have top priority. By means of a scoring system, cloud service providers are rated with one to five stars, expressing the degree of fulfillment of aforementioned criteria and therefore the trustworthiness. In the near future EuroCloud Star Audit will be expanded to the other two subdomains of cloud computing, namely Platform-as-a-Service (Paas) and Infrastructure-as-a-Service (IaaS), to enable a more complete rating of cloud service providers [32].

The second trust-building measure that should help choose the right cloud service provider is standards compliance. The International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC) created a series of information security standards, namely the 27000-series. It provides best practice recommendations on information security management, risks and control within the context of an overall Information Security Management System (ISMS). The series is applicable to all types and sizes of organizations and, most importantly, for cloud service providers. Among other topics it covers privacy, confidentiality and IT or technical security issues. The standards series includes ISO/IEC 27001, a standard that specifies requirements for establishing, implementing, operating, monitoring, reviewing, maintaining and improving a documented Information Security Management System within the context of the organization's overall business risks. It specifies requirements for the implementation of security controls customized to the needs of individual organizations or parts thereof. It is designed to ensure the selection of adequate and proportionate security controls that protect information assets and give confidence to interested parties. The succeeding standards ISO/IEC 27003, 27004, 27005 and 27006 all refer to the requirements defined in 27001. ISO/IEC 72003 focuses on the critical aspects needed for successful design and implementation of an ISMS. ISO/IEC 27004 provides guidance on the development and use of measures and measurement in order to assess the effectiveness of an implemented ISMS. ISO/IEC 27005 specifies guidelines for information security risk management and ISO/IEC 27006 specifies requirements and guidance for bodies providing audit and certification of an ISMS and is primarily intended to support the accreditation of certification bodies providing ISMS certification [33]. By implementing an ISO/IEC 27001 information security management system, the organization adopts a comprehensive and systematic approach to the security of the process control systems and can therefore be formally audited and certified compliant with the standard.

The third trust-building measure that should help choose the right cloud service provider is IT service quality as defined in the IT Infrastructure Library (ITIL) framework. It is independent of manufacturers, and describes systematic procedures for the strategic development, design, introduction, transition, operation and improvement of IT services. It closely follows ISO/IEC 20000, which provides a formal and universal standard for organizations seeking to have their service management capabilities audited and certified. ITIL version 3, passed in June 2007, consists of five books: Service strategy, service design, service transition, service operation and continual service improvement. Cloud service providers that have aligned their services to the ITIL framework can increase their trustworthiness not only, but mainly because of three ITIL building blocks:

- Information Security Management (ISM)
- Availability Management
- Access Management

ISM ensures most of the information security key concepts: Confidentiality, integrity and availability of an organization's assets, information, data and IT services. Information security is aligned with business security and ISM ensures that information security is effectively managed in all service management processes, activities, etc. The ISM process should be a focal point for all IT security issues and should increase awareness of the need for security within all IT services. A main task of ISM is to produce, maintain and enforce the information security policy.

Availability Management focuses and manages all availability-related issues and is responsible for defining, analyzing, planning, measuring and improving all aspects of the availability of IT services. It ensures that the IT infrastructure and processes, tools, roles etc. are appropriate for the agreed service level targets for availability. This process thus secures the level of availability delivered in all services is matched to, or exceeds the current and future agreed needs of the customers in a cost-effective manner. Availability Management is important because availability and reliability are highly visible to the customers and can directly influence customer satisfaction and the service provider's reputation.

Access Management deals with protecting the confidentiality, integrity and availability of the organization's data and intellectual property. It achieves this by ensuring that only authorized users are able to access or modify the service assets. It provides the right for users to use a service or group of services, while preventing access to non-authorized users. It may also be needed for regulatory compliance reasons. Technologically, Access Management is usually executed by means of directory services [14][34].

All of the three suggested trust-building measures have one thing in common: They prove through examination of a trusted third party that the cloud service provider operates with the necessary care and accuracy required by the presented certifications, standards, frameworks and grants compliance. The willingness of the provider to do so creates transparency for the cloud users and a chance to make a well-informed decision.

G. *Transparency*

As learned, trust is always a subjective matter, which gives transparency requirements for trust a soft and elastic touch. Transparency has multiple facets though. Trust through transparency can be induced by very simple means such as a web interface design or by more sophisticated means such as a conglomeration of technological factors.

In [11], a framework of four trust-inducing features is proposed by taking existing relevant studies on enhancing online trust by web interface design and using them as dimensions of the framework. The four dimensions are graphic design, structure design, content design and social-cue design. Graphic design refers to the graphical design factors on the web site that normally give consumers a first impression. Structure design defines the overall organization and accessibility of displayed information on the web site. Content design refers to the informational components that can be included on the web site, either textual or graphical. Social-cue design relates to embedding social cues, such as face-to-face interaction and social presence into web interface via different communication media.

Compared to a trust-inducing web interface design, transparency as add on to technological security mechanisms has much clearer and more precise requirements. Contradicting the often-used principle of security by obscurity, T. Weichert demands security by transparency [31]. He sets up multiple factors on how to reach this goal:

- State of the art measures
- Access restricted to entitled users
- Differentiated access management
- Encryption capabilities
- Anonymization tools
- Adequate separation of data by isolating
- Client-side application security
- Documented data privacy management

His statement is simple to understand: The more of these factors are in place, the higher the transparency and therefore security for cloud service customers will be.

VI. CONCLUSION AND FUTURE WORK

Cloud computing services will grow further, regardless of whether a cloud service provider sells services at a low level of abstraction as IaaS, at the medium level as PaaS or at the top level as SaaS. Trust and security go hand in hand - one might even go as far as saying one induces the other.

This paper presented a state of the art overview of the role of trust in cloud computing. Explaining and mapping offline trust to online trust, we showed that the concept of trust does also exist and even plays a vital role in the online world. Trust and security are an integral part of cloud computing and essential for its adoption and growth.

Our main contribution is showing multiple ways to improve online trust and security by leveraging and combining as many existing technology and trust building measures as possible, and by that, minimizing concerns of potential or existing cloud service users. In our opinion, the bottom line of this state of the art overview is, that trust in cloud computing can indeed be improved by means of technology.

A. Limitations

The paper did provide several existing approaches to the issue of insufficient trust and security in cloud computing. However, there are several limitations that have to be acknowledged. The paper did not examine infrastructure issues such as data transfer bottlenecks and performance unpredictability. Computing, storage and networking must all focus on horizontal scalability of virtualized resources rather than on single node performance. Infrastructure in all areas has to be improved, not only in respect to trust and security, but also in respect bandwidth and cost. Furthermore, the paper only highlighted a fractional amount of available security and trust enhancing mechanisms, which we found most important. There are a large number of other efficient mechanisms, standards and an even larger number under investigation in research and development.

B. Future Research and Recommendations

This paper's examples contribute to the ongoing effort of minimizing the challenges regarding trust and security in cloud computing. What still remains is the issue that users have to trust the presented technology, certifications, standards and finally the cloud service provider itself.

Even though trust per definition remains the willingness of a party to be vulnerable to the actions of another party, many unsolved technical issues still exist and many solutions can be improved in order to reduce this inevitable residual risk.

Future research on this topic should include the simplification of cloud security models, for example by standardizing and leveraging protocols, such as the Open Authorization Protocol (OAuth) and the Security Assertion Markup Language (SAML). With the vision of Inter-Cloud-Computing in mind, which introduces an additional management layer above conventional cloud computing systems [35] to reach greater sustainability and availability, large IT companies have to work together more intensely in taskforces, alliances and foundations to push towards this common goal.

ACKNOWLEDGMENT

The authors gratefully acknowledge support from grant 17103X10 from the German federal ministry of education and research.

REFERENCES

[1] C. Weinhardt, A. Anandasivam, B. Blau, N. Borissov, T. Meinl, W. Michalk, and J. Stößer, "Cloud Computing – A classification, business models, and research directions," *Business & Information Systems Engineering*, 5, pp. 391–399, 2009.

[2] C. Baun, M. Kunze, T. Kurze, and V. Mauch, "Private Cloud-Infrastrukturen und Cloud-Plattformen," *Informatik Spektrum*, vol. 34, no. 3, pp. 242–254, 2011.

[3] Cloud Security Alliance, (2009) "Security guidance for critical areas of focus in Cloud Computing," [Online]. Available: https://cloudsecurityalliance.org/wp-content/themes/csa/guidance-download-box.php [retrieved: April, 2012]

[4] A. Weiss, "Computing in the clouds," *networker*, vol. 11, no. 4, pp. 16–25, 2007.

[5] F. Kamoun, "Virtualizing the datacenter without compromising server performance," *Ubiquity*, vol. 2009, no. August, p. 2, 2009.

[6] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *ACM SIGCOMM Comput. Commun.*, vol. 39, no. 1, p. 68, 2009.

[7] B. Hayes, "Cloud Computing," *Comm. ACM*, vol. 51, no. 7, p. 9, 2008.

[8] S. Pearson, "Taking account of privacy when designing cloud computing services," in *Proceedings of the 2009 ICSE Workshop on Software Engineering Challenges of Cloud Computing (CLOUD '09)*, pp. 44–52.

[9] M. Armbrust, I. Stoica, M. Zaharia, A. Fox, R. Griffith, A. D. Joseph, et al., "A view of cloud computing," *Comm. ACM*, vol. 53, no. 4, p. 50, 2010.

[10] J. Bosch, S. Clarke, M. Mowbray, and S. Pearson, "A client-based privacy manager for cloud computing," in *COMSWARE '09 Proceedings of the Fourth International ICST Conference on COMmunication System softWAre and middlewaRE*, p. 1, 2009.

[11] Y. D. Wang and H.H. Emurian, "An overview of online trust: Concepts, elements, and implications," *Computers in Human Behavior*, vol. 21, no. 1, pp. 105–125, 2005.

[12] E. H. Erikson, "Childhood and society" (2nd ed.), New York: W.W. Norton, 1963.

[13] M. Deutsch, "Cooperation and trust: Some theoretical notes, " *Nebraska Symposium on Motivation*, 10, pp. 275–318, 1962.

[14] J. B. Rotter, "A new scale for the measurement of interpersonal trust," *J of Personality*, vol. 35, no. 4, pp. 651–665, 1967.

[15] B. Uzzi, "Social structure and competition in interfirm networks: The paradox of embeddedness," *Administrative Science Quarterly*, vol. 42, no. 1, pp. 35–67, 1997.

[16] C. Moorman, R. Deshpande, and G. Zaltman, "Factors affecting trust in market research relationships," *J of Marketing*, vol. 57, no. 1, pp. 81–101, 1993.

[17] Y. Tan and W. Thoen, "Toward a generic model of trust for electronic commerce," *International J of Electronic Commerce*, vol. 5, no. 2 (Winter, 2000/2001), pp. 61-74, 2001.

[18] R. C. Mayer, J. H. Davis, and F. D. Schoorman, "An integrative model of organizational trust," Academy of Management Review, vol. 20, no. 3, pp. 709–734, 1995.

[19] T. H. Noor and Q. Z. Sheng, "Trust as a service: A framework for trust management in cloud environments," pp. 314–321, 2011.

[20] G. Zhao, C. Rong, M. G. Jaatun, and F. E. Sandnes, "Reference deployment models for eliminating user concerns on cloud security," *J of Supercomputing*, 2010.

[21] P. Laue and O. Stiemerling, "Identitäts- und Zugriffsmanagement für Cloud Computing Anwendungen," *Datenschutz und Datensicherheit*, vol. 34, no. 10, pp. 692–697, 2010.

[22] N. Sonehara, I. Echizen, and S. Wohlgemuth, "Isolation in cloud computing and privacy-enhancing technologies," *Business & Information Systems Engineering*, vol. 3, no. 3, pp. 155–162, 2011.

[23] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai network; a platform for high-performance internet applications," *SIGOPS Oper. Syst.*, vol. 44, no. 3, p. 2, 2010.

[24] A. Undheim, A. Chilwan, and P. Heegaard, "Differentiated availability in cloud computing SLAs," in *Proceedings of the 2011 IEEE/ACM 12ith International Conference on Grid Computing (GRID '11)*, pp. 129–136.

[25] P. Gill, N. Jain, and N. Nagappan, "Understanding network failures in data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 4, pp. 350-361, 2011.

[26] J. M. Hellerstein, S. Chaudhuri, and M. Rosenblum, K. V. Vishwanath, N. Nagappan, "Characterizing cloud computing hardware reliability," in *Proceedings of the 1st ACM symposium on Cloud Computing (SoCC '10)*, p. 193-204, 2010.

[27] W. Hao, I. Yen, and B. Thuraisingham, "Dynamic service and data migration in the clouds," in *Computer Software and Applications Conference, COMPSAC '09. 33rd Annual IEEE International*, pp. 134–139, 2009.

[28] T. Hirofuchi, H. Ogawa, H. Nakada, S. Itoh, and S. Sekiguchi, "A live storage migration mechanism over WAN for relocatable virtual machine services on clouds," in *Proceedings of the 2009 9th IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID '09)*, pp. 460–465.

[29] F. Kerschbaum, "Secure and sustainable benchmarking in clouds," *Business & Information Systems Engineering*, vol. 3, no. 3, pp. 135–143, 2011.

[30] S: Paulus, "Standards für trusted clouds," *Datenschutz und Datensicherheit*, vol. 35, no. 5, pp. 317–321, 2011.

[31] T. Weichert, "Cloud Computing und Datenschutz," *Datenschutz und Datensicherheit*, vol. 34, no. 10, pp. 679–687, 2010.

[32] R. Giebichenstein and A. Weiss, "Zertifizierte Cloud durch das EuroCloud Star Audit SaaS," *Datenschutz und Datensicherheit*, vol. 35, no. 5, pp. 338–342, 2011.

[33] International Organization for Standardization at http://www.iso.org [retrieved: April, 2012]

[34] Materna Information & Communications, (2012), ITIL Version 3 Pocket Guide [Online]. Available: http://www.materna.de/cae/servlet/contentblob/11600/publicationFile/2465/Pocketbrosch%C3%BCre ITIL%C2%AE Version 3.pdf [retrieved: April, 2012]

[35] T. Aoyama and H. Sakai, "Inter-cloud computing," *Business & Information Systems Engineering*, vol. 3, no. 3, pp. 173–177, 2011.

Application of Machine Learning Algorithms to an online Recruitment System

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis
Computer Engineering and Informatics Department
University of Patras
Patras, Greece
faliagka@ceid.upatras.gr, ramantas@ceid.upatras.gr, tsak@cti.gr

Giannis Tzimas
Department of Applied Informatics in Management & Economy, Faculty of Management and Economics
Technological Educational Institute of Messolonghi
Messolonghi, Greece
tzimas@cti.gr

***Abstract*— In this work, we present a novel approach for evaluating job applicants in online recruitment systems, leveraging machine learning algorithms to solve the candidate ranking problem. An application of our approach is implemented in the form of a prototype system, whose functionality is showcased and evaluated in a real-world recruitment scenario. The proposed system extracts a set of objective criteria from the applicants' LinkedIn profile, and infers their personality characteristics using linguistic analysis on their blog posts. Our system was found to perform consistently compared to human recruiters; thus, it can be trusted for the automation of applicant ranking and personality mining.**

***Keywords** - e-recruitment; personality mining; recommendation systems; data mining.*

I. INTRODUCTION

The rapid development of modern Information and Communication Technologies (ICTs) in the past few years has resulted in an increasing number of people turning to the web for job seeking and career development. A lot of companies use online knowledge management systems to hire employees, exploiting the advantages of the World Wide Web. These are termed e-recruitment systems and automate the process of publishing positions and receiving CVs. E-recruitment systems have seen an explosive expansion in the past few years [1], allowing Human Resources (HR) agencies to target a very wide audience at a small cost. This situation might be overwhelming to HR agencies that need to allocate human resources for manually assessing the candidate resumes and evaluating the applicants' suitability for the positions at hand. Automating the process of analyzing the applicant profiles to determine the ones that fit the position's specifications could lead to an increased efficiency. For example, SAT telecom reported 44% cost savings and a drop in the average time needed to fill a vacancy from 70 to 37 days [2] after deploying an e-recruitment system.

Several e-recruitment systems have been proposed with an objective to speed-up the recruitment process, leading to a better overall user experience. E-Gen system [3] performs analysis and categorization of unstructured job offers (i.e., in the form of unstructured text documents) as well as analysis and relevance ranking of candidates. CommOn framework [4] applies Semantic Web technologies in the field of Human Resources Management. In this framework, the candidate's personality traits, determined through an online questionnaire which is filled-in by the candidate, are considered for recruitment. In order to match applicants with job positions these systems typically combine techniques from classical IR and recommender systems, such as relevance feedback [3], semantic matching [5] and Analytic Hierarchy Process [6]. Another approach proposed in [7] uses NLP technology to automatically represent CVs in a standard modeling language. These methods, although useful, suffer from the discrepancies associated with inconsistent CV formats, structure and contextual information. What's more they are unable to evaluate some secondary characteristics associated with CVs, such as style and coherence, which are very important in CV evaluation.

In this work, we propose the application of supervised learning algorithms in automated e-recruitment systems, to solve the candidate ranking problem. What's more, we have implemented and tested an integrated company oriented e-recruitment system that automates the candidate pre-screening and ranking process. In the proposed system, the applicants' evaluation is based on a predefined set of objective criteria, which are directly extracted from the applicant's LinkedIn profile. What's more, the candidate's personality characteristics, which are automatically extracted from his social presence [8], are taken into account in his evaluation. Our objective is to limit interviewing and background investigation of applicants solely to the top candidates identified from the system, so as to increase the efficiency of the recruitment process. The system is designed with the aim of being integrated with the companies' Human Resource Management infrastructure, assisting and not replacing the recruiters in their decision-making process.

The rest of this work is organized as follows. In Section II, we present an overview of the proposed e-recruitment system. In Section III, a personality mining scheme is proposed, to extract the applicant's personality traits from textual data available for the candidate in the web. In

Section IV the supervised learning algorithms used to rank the candidates are detailed, and, in Section V, we present a set of experimental results that showcase the effectiveness of our system in a real-world recruitment scenario. Finally, the proposed system was implemented in the form of a web application, whose design and prototype implementation is presented in Section VI.

II. SYSTEM OVERVIEW

In this work, we have implemented an integrated company-oriented e-recruitment system that automates the candidate evaluation and pre-screening process. Its objective is to calculate the applicant's relevance scores, which reflect how well their profile fits the positions' specifications. In this section, we present an overview of the proposed system architecture and candidate ranking scheme.

A. *Architecture*

The proposed e-recruitment system implements automated candidate ranking based on a set of credible criteria, which will be easy for companies to integrate with their existing Human Resources Management infrastructure. In this study we focus on 4 complementary selection criteria, namely: Education (in years of formal academic training), Work Experience, Loyalty (average number of years spent per job) and Extraversion. The system architecture, which is shown in Fig. 1, consists of the following components:

- *Job Application* module: It implements the input forms that allow the candidates to apply for a job position. The candidate is given the option to log into our system using his LinkedIn account credentials, which allows the system to automatically extract all objective selection criteria directly from the user's LinkedIn profile.
- *Personality mining* module: If the candidate's blog URL is provided, it applies linguistic analysis to his blog posts to derive features reflecting the author's personality.
- *Applicant Grading* module: It combines the candidate's selection criteria to derive the candidate's relevance score for the applied position. The grading function is derived through supervised learning algorithms.

Each applicant's qualifications, as well as his relevance score, are stored in the system's database. At the end of the recruitment process, the top candidates are called to participate in the interview process. It must be noted here that during the job application process, the applicant is not required to manually enter information or participate in time-consuming personality tests. Thus, the user friendliness and the practicality of the system are maintained.

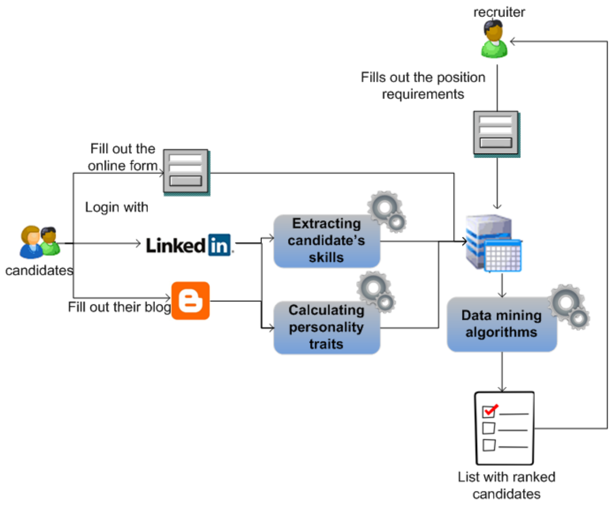

recruiter
Fills out the position requirements
Fill out the online form
Login with
Linked in
candidates
Fill out their blog
Extracting candidate's skills
Calculating personality traits
Data mining algorithms
List with ranked candidates

Figure 1. System Architecture

B. *Candidate Ranking*

The increasing number of submitted CVs may overwhelm HR departments, which typically perform manual evaluation of job applications. Automated candidate ranking systems, that have been proposed to speed-up the recruitment process typically require a model of the HR department's decision making process, as well as a careful parameterization by the department's expert recruiters. This is a complex and error-prone procedure, which must be repeated each time the selection criteria change. The proposed system leverages machine learning algorithms to automatically build the applicant ranking models. This approach requires sufficient training data as an input, which consist of previous candidate selection decisions. Methods that learn how to combine predefined features for ranking by means of supervised learning algorithms are called "learning-to-rank" methods. In recent years, learning to rank has become a hot research direction in information retrieval [9], but it can also be applied in many real-world ranking problems.

In Fig. 2, the typical "learning to rank" process is shown. A training set is used that consists of past candidate applications represented by feature vectors, denoted as $x_i^{(k)}$, along with an expert recruiter's judgment of the candidate's relevance score, denoted as y_i. Candidate's features can be assessed either on a numerical scale (e.g., years of work experience) or with a Boolean variable, which represents whether the candidate reports a certain skill or not in his LinkedIn profile. The training set is fed to a learning algorithm which constructs the ranking model, such that its output predicts the recruiter's judgment when given the candidate's feature vector as an input. In the test phase, the learned model is applied to sort a set of candidate applications, and return the final ranked list of candidates. Many learning-to-rank algorithms can fit in the abovementioned process, and each one models the process of learning to rank in a different way.

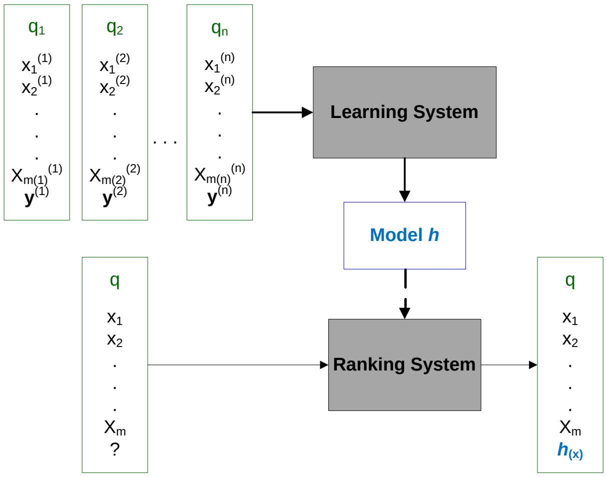


Figure 2. The “learning to rank” process

III. PERSONALITY MINING

The applicants’ personality traits are critical for their selection in many job positions, but are usually overlooked in existing e-recruitment systems. Typically, candidates’ personality is assessed during the interview stage, which is reserved to the candidates that passed the pre-screening phase. However, gathering some preliminary data for the candidate’s personality in the pre-screening phase is considered valuable, and such information is often obtained through web searches. In the Web 2.0 era, there are large amounts of textual data for millions web users, that have been shown to be reliable predictors of user’s personality. The proposed system automates the task of personality mining using text analysis, an approach proposed in [8].

Previous works have shown that by applying linguistic analysis to blog posts, the author’s personality traits can be derived, [10] as well as his mood and emotions [11]. The text analysis in these works is performed with LIWC (Linguistic Inquiry and Word Count) system, which extracts linguistic features that act as markers of the author’s personality. LIWC uses a dictionary of word stems classified in certain psycholinguistic semantic and syntactic word categories. It analyzes written text samples by counting the relative frequencies of words that fall in each word category. Pennebaker and King have found significant correlations between these frequency counts and the author’s personality traits [12] as measured by the Big-Five personality dimensions.

In this work, we focus on the extraversion personality trait, due to its importance in candidate selection. Extraversion is a crucial personality characteristic in positions that interact with customers, while social skills are important for team work. It has been shown that extraversion is adequately reflected through language use in written speech and it is possible to be discriminated through text analysis. Specifically, the emotional positivity and social orientation of candidates, both directly extracted from LIWC frequencies, can act as predictors of extroversion trait [8].

In this work, an expert recruiter has assigned extraversion scores to each of 100 job applicants with personal blogs, which were part of a large-scale recruitment scenario (see Section V for a detailed description of the scenario). The recruiter’s scores were used to train a regression model, which predicts the candidates’ extraversion from their LIWC scores in the {posemo, negemo, social} categories. In what follows, a linear regression model was selected as a predictor of the extraversion score E, as proposed in [13], due to its good accuracy and low complexity. Equation (1) corresponds to the linear model that minimizes the Mean Square Error between actual values assigned by the recruiter and predicted scores output by the model:

$$E = S + 1.335 * P - 2.250 * N \,, \qquad (1)$$

where S is the frequency of social words (such as friend, buddy, coworker) returned from LIWC, P the frequency of positive emotion works and N the frequency of negative emotion words.

IV. LEARNING TO RANK ALGORITHMS

In this work, we leverage machine learning techniques to solve the candidate ranking problem in e-recruitment systems. In the candidate ranking problem, a scoring function h(x) outputs the candidate relevance score, which reflects how well a candidate profile fits the requirements of a given job position. As the relevance score is a continuous variable, the candidate ranking problem can be reduced to a regression problem where the candidate scoring function must be learned using supervised learning techniques. Then the system outputs the final ranked list by applying the learned function to sort the candidates. The score function h(x) derives the candidate’s relevance degree y_i from the values of his feature vector x_i. In this work, the feature vector x_i consists of a set of m attributes {a_1, …, a_m} that correspond to the candidate’s selection criteria. These can be either continuous variables (representing a candidate’s feature assessed on numerical scale) or Boolean variables (declaring whether he has a desired skill or not). The true scoring function is usually unknown and an approximation is learned from the training set D. In the proposed system the training set consists of a set of N previous candidate selection examples, given as an input to the system:

$$D = \{(x_i, y_i) \mid x_i \in R^m, y_i \in R\}_{i=1}^{N} \,. \qquad (2)$$

In what follows, we present a set of representative “learning to rank” algorithms [9] that map the training set D of previous recruiting decisions to a regression model that serves as a predictor of future recruiting decisions.

1) *Linear Regression*

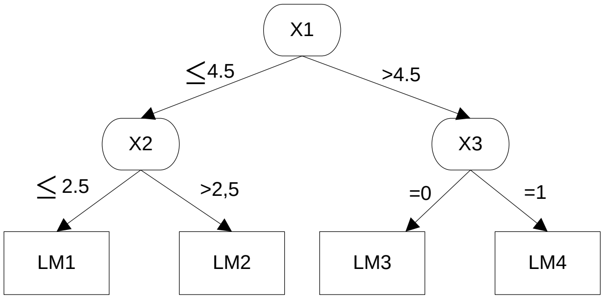


Figure 3. M5 Model tree

In linear regression, the relevance score y_i of the i^{th} candidate is predicted as a linear function of the selection criteria, which comprise the candidate's feature vector x_i plus noise *e* (regression error):

$$y_i = w^T x_i + e. \quad (3)$$

The linear regression algorithm finds the optimal parameter vector w that minimizes the regression error.

2) Regression Tree

When selection criteria interact in complex and non-linear ways, linear regression that constructs a linear prediction formula for all data space is not an appropriate model. Regression trees can be a viable alternative, as they recursively partition the predictor space using a divide and conquer approach. They have the same structure as propositional decision trees; internal nodes contain tests and leaves contain predictions for the class value (see Fig. 3). In our experiments, we use an M5' model tree and a REPTree regression tree.

3) Support Vector Regression

Support Vector Machines (SVMs) are a set of related methods for supervised learning, applicable to both classification and regression problems. The power of SVMs comes from the kernel representation, which allows a non-linear mapping of input space to a higher dimensional feature space. The objective of Support Vector Regression is to find a function *f* that minimizes the expected error – i.e., the integral of a certain loss function – according to the unknown probability distribution of the data. This minimizes the empirical risk that the estimated function differs from the original (yet unknown) one. Assuming N data points and a Kernel K, the support vectors and the support values of the solution define the following regression function:

$$f(x) = \sum_{i=1}^{N} a_i K(x, x_i) + b \mid b, a_i \in R. \quad (4)$$

V. Experimental Evaluation

The proposed system was tested in a real-world recruitment scenario, to evaluate its effectiveness in ranking job applicants. The system's performance evaluation is based on how effective it is in assigning consistent relevance scores to the candidates, compared to the ones assigned by human recruiters.

A. Data Collection

In the recruitment scenario used in our tests, we compiled a corpus of 100 applicants with a LinkedIn account and a personal blog, as these are key requirements of the proposed system. The applicants were selected randomly via Google blog search API with the sole requirement of having a technical background, as indicated by the blog metadata (list of interests), as well as a LinkedIn profile. Our corpus of job applicants was formed by choosing the first 100 blogs returned from the profile search API that fulfilled our preconditions. We also collected three representative technical positions announced by an unnamed IT company with different requirements, i.e., a sales engineering position, a junior programmer position and a senior programmer position.

The sales engineering position favors a high degree of extraversion, while experience is the most important feature for senior programmers. Junior programmers are mainly judged by loyalty (because a company would not invest in training an individual prone to changing positions frequently) as well as education. What's more, each position has its own desired set of skills, which are matched with the skillset reported by each user at his LinkedIn profile. Specifically, the junior position requires programming skills in C++ or Java development languages, while the senior position requires a 5-year experience in J2EE technologies. The use of different requirements per position is expected to test the ability of our system to match candidate's profiles with the appropriate job position.

B. Experimental Results

In our experiments, we assume that each applicant in the corpus has applied for all three available job positions. For each job position, applicants were ranked according to their suitability for the job position both by the system (automated ranking) and by an expert recruiter. Human recruiters had access to the same information as the system, i.e., the candidate's blog and LinkedIn profile. It must be

TABLE I. Correlation Coefficients For Applicants' Relevance Scores Vs Different Machine Learning Models

Correlation coefficient	LR	M5' Tree	REP Tree	SVR, poly	SVR, PUK
Sales engineer	0.74	0.81	0.81	0.61	0.81
Junior programmer	0.79	0.85	0.84	0.81	0.84

Senior programmer	0.64	0.63	0.68	0.62	0.73

noted though that despite the fact that the selection criteria are known to the system, the recruiter's interpretation of the data and the exact decision-making process is unknown and must be learned.

In our first experiment, we use Weka [14] to evaluate the learning-to-rank models. Specifically, we test the correlation of the scores output from the system (i.e., model predictions) with the actual scores assigned by the recruiters, using the Pearson's correlation coefficient metric. Table I shows the correlation coefficients for 4 different machine learning models, namely: Linear Regression (LR), M5' model tree (M5'), REPTree decision tree (REP), and Support Vector Regression (SVR) with two non-linear kernels (i.e., polynomial kernel and PUK universal kernel). It can be seen that the Tree models and the SVR model with a PUK kernel produce the best results. On the other hand, Linear Regression performs poorly, suggesting that the selection criteria are not linearly separable. It must be noted here that all values are averages, obtained with the 10-fold cross validation technique.

It can be seen in Table I that the consistency of the system's scores is highly dependent on the nature of the offered positions. For the sales position, the recruiter's judgment is dominated by the highly subjective extraversion score, thus increasing the uncertainty of the overall relevance score. Still, the system was able to achieve a correlation coefficient of up to 0.81, depending on the regression model used. On the other hand the selection of junior programmer candidates is based on more objective criteria such as loyalty and education, thus resulting in a slightly higher correlation coefficient, up to 0.85. Finally, the senior programmer's position exhibited the lowest consistency, with a Pearson's correlation of up to 0.73. This can be attributed to the high complexity of building a regression model for a senior position, which typically requires domain-specific experience and specific qualifications.

In our second experiment, we evaluate the effectiveness of the personality mining scheme, presented in Section III. As mentioned earlier, our system exploits textual data from the candidate's blog to predict his extraversion score, as determined by an expert recruiter who had access to the same blog posts. The extraversion score is predicted by training a regression model to the extroversion scores assigned from the recruiter to each of the 100 candidates. In this experiment we use Weka to test the effectiveness of 4 different regression models, compiling a table (Table II) with the Pearson's correlation coefficients and relative errors between system's and recruiter's scores. It must be noted that regression models try to replicate the actual scalar values associated by the recruiter, which is a hard problem. Nevertheless, a significant correlation was found, with a Pearson's coefficient of up to 0.65.

TABLE II. CORRELATION COEFFICIENTS AND RELATIVE ERRORS FOR APPLICANT'S EXTRAVERSION SCORE VS MACHINE LEARNING MODELS

Correlation coefficient	**LR**	**M5' Tree**	**REP Tree**	**SVR, poly**	**SVR, PUK**
Pearson's Coefficient	0.63	0.63	0.65	0.28	0.65
Relative error	25.3%	25.3%	22.5%	57.4%	23.1%

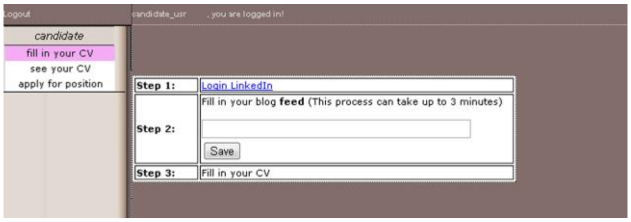
VI. PROTOTYPE IMPLEMENTATION

The proposed e-recruitment system was fully implemented as a web application, in the Microsoft .Net development environment. In this section we will present the main application screens and discuss our design decisions and system implementation. The system is divided in the recruiter's side and the user's side.

A. *Job application process (user's side)*

Job applicants are given the option to authenticate using their LinkedIn account credentials (see Fig. 4) to apply for one or more of the available job positions. This allows the system to automatically extract the selection criteria required for candidate pre-screening from the applicants' LinkedIn profile, so the user experience is streamlined. Users are authorized with LinkedIn API, which uses OAuth [15] as its authentication protocol. After successful user authentication, an OAuth token is returned to our system which allows retrieving information from the candidate's private LinkedIn profile. It must be noted here that the system does not have direct access to the candidate's account credentials, which could be regarded as a security risk. Users without a LinkedIn profile are given the option to enter the required information manually.

As part of the job application process, the candidate is asked to fill-in the feed URI of his personal blog. This allows our system to syndicate the blog content and calculate the extraversion score with the personality mining technique presented in Section III. Blog posts are input to the TreeTagger tool [16] for lexical analysis and lemmatization. Then, using the LIWC dictionary which is distributed as part of the LIWC tool, our system classifies the canonical form of words output from TreeTagger in one of the word categories of interest (i.e., positive emotion,



Step 1:	Login LinkedIn
Step 2:	Fill in your blog **feed** (This process can take up to 3 minutes)
Step 3:	Fill in your CV

Figure 4. Job application process

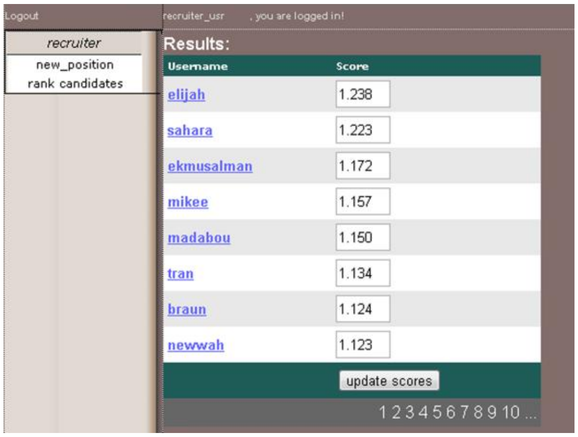
negative emotion and social words) and calculates the LIWC scores. Finally, the system estimates the applicant's extraversion score.

B. *Recruitment process (recruiter's side)*

After authenticating with their account credentials, recruiters have access to the recruitment module, which gives them rights to post new job positions and evaluate job applicants. In the "rank candidates" menu, the recruiter is presented with a list of all available job positions and the candidates that have applied for each one of them. Upon the recruiter's request, the system estimates applicants' relevance scores and ranks them accordingly. This is achieved by calling the corresponding Weka classifier, via calls to the API provided by Weka. The recruiter can modify the candidate ranking, by assigning his own relevance scores to the candidates, as shown in Fig. 5. This will improve the future performance of the system, as the recruiter's suggestions are incorporated in the system's training set and the ranking model is updated. It must be noted here that the ranking model is initialized as a simple linear combination of the selection criteria, until sufficient input is provided from the recruiters to build a training set.

VII. CONCLUSIONS

In this paper, we have presented a novel approach for ranking job applicants in online recruitment systems. The proposed scheme relies on objective criteria extracted from the applicants' LinkedIn profile and subjective criteria extracted from their social presence, to estimate applicants' relevance scores and infer their personality traits. Candidate ranking is based on machine learning algorithms that learn the scoring function based on training data provided by human recruiters. An integrated company oriented e-recruitment system was implemented based on the proposed scheme. Our system was employed in a large-scale recruitment scenario, which included three different offered positions and 100 job applicants. The application of our approach revealed that it is effective in identifying the job applicants' extraversion and ranking them accordingly.



recruiter_usr , you are logged in!

Results:

Username	Score
elijah	1.238
sahara	1.223
ekmusalman	1.172
mikee	1.157
madabou	1.150
tran	1.134
braun	1.124
newwah	1.123

Figure 5. Candidate ranking results

REFERENCES

[1] P. De Meo, G. Quattrone, G. Terracina and D. Ursino, "An XML-Based Multiagent System for Supporting Online Recruitment Services," Systems, Man and Cybernetics, Part A: Systems and Humans, vol. 37, July. 2007, pp. 464 – 480.

[2] S. Pande, "E-recruitment creates order out of chaos at SAT Telecom: System cuts costs and improves efficiency", Human Resource Management International Digest, Vol. 19, 2011 pp. 21–23.

[3] R. Kessler, J. Torres-Moreno and M. El-Beze, "E-Gen: automatic job offer processing system for human resources". Proc. of MICAI'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 985-995.

[4] V. Radevski and F. Trichet, "Ontology-Based Systems Dedicated to Human Resources Management: An Application in e-Recruitment," On the Move to Meaningful Internet Systems, vol. 4278, 2006, pp. 1068–1077.

[5] M. Mochol, H. Wache, and L. Nixon, "Improving the Accuracy of Job Search with Semantic Techniques", Business Information Systems, vol. 4439, 2007, pp. 301-313.

[6] E. Faliagka, K. Ramantas, A. Tsakalidis, M. Viennas, E. Kafeza and G. Tzimas, "An Integrated e-Recruitment System for CV Ranking based on AHP," Proc. of WEBIST 2011, May. 2011, pp. 147-150.

[7] S. Amdouni and W. Ben Abdessalem Karaa, "Web-based recruiting", Proc. Of International Conference on Computer Systems and Applications (AICCSA), 2010, pp. 1-7.

[8] E. Faliagka, L. Kozanidis, S. Stamou, A. Tsakalidis and G. Tzimas, "Personality Mining System for Automated Applicant Ranking in Online Recruitment Systems,"Proc. of ICWE 2011, Springer-Verlag, Berlin, Heidelberg, June. 2011, pp. 379-382.

[9] T. Liu, "Learning to Rank for Information Retrieval," Foundations and Trends in Information Retrieval, vol. 3, March 2009, pp. 225-331

[10] J.A. Gill, S. Nowson and J. Oberlander, "What are they blogging about? Personality, topic, and motivation in blogs", Proc. of AAAI ICWSM. 2009

[11] G. Mishne, "Experiments with mood classification in blog posts", Proc. of 1st Workshop on Stylistic Analysis Of Text For Information Access Style 2005. 2005

[12] J.W. Pennebaker and L. King, "Linguistic Styles: Language Use as an Individual Difference," Journal of Personality and Social Psychology, vol. 77, 1999, pp. 1296–1312.

[13] F. Mairesse, M.A. Walker, M.R. Mehl and R.K. Moore, "Using linguistic cues for the automatic recognition of personality in conversation and text," Journal of Artificial Intelligence Research, vol. 30, 2007, pp. 457-500.

[14] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. Witten, "The WEKA data mining software: an update," SIGKDD Explorer, News, vol. 11, 2009, pp. 10-18.

[15] E. Hammer-Lahav and D. Recordon, "The OAuth 1.0 Protocol", http://tools.ietf.org/html/draft-hammer-oauth-10, February 2010.

[16] H. Schmid, "Improvements In Part-of-Speech Tagging With an Application To German", Proc. of ACL SIGDAT, 1995, pp. 47-50.

An Approach for Performance Test Artefact Generation for Multiple Technologies from MARTE-Annotated Workflows

Antonio García-Domínguez and Inmaculada Medina-Bulo
Department of Computer Languages and Systems
University of Cádiz
Cádiz, Spain
{antonio.garciadominguez, inmaculada.medina}@uca.es

Mariano Marcos-Bárcena
Department of Industrial Design and Mechanical Engineering
University of Cádiz
Cádiz, Spain
mariano.marcos@uca.es

***Abstract*—Obtaining the expected performance from a workflow would be easier if every task included its own specifications. However, normally only global performance requirements are provided, forcing designers to infer individual requirements by hand. In previous work we presented two algorithms that automatically inferred local performance constraints in Unified Modelling Language activity diagrams annotated with the Modelling and Analysis of Real-Time and Embedded Systems profile. In this work, we present an approach to use these annotations to generate performance test cases for multiple technologies, linking a performance model and an implementation model with a weaving model. We describe how it can be applied to Java code and to Web Service compositions, using existing open source technologies and discussing the challenges involved. The resulting processes follow a meet-in-the-middle approach, allowing the user to write their software according to their needs.**

***Keywords*-software performance; Web Services; MARTE; model weaving; model driven engineering.**

I. Introduction

Software needs to meet both functional and non-functional requirements. Performance requirements are among the most commonly used non-functional requirements, and in some contexts they can be just as important as functional requirements. In addition to soft and hard real-time systems, Service Oriented Architectures (SOAs) must be considered as well. Within SOAs, it is common practice to sign Service Level Agreement (SLAs) with external services, to compensate consumers in case of problems. It is also quite common to create "service compositions", which are services that integrate several lower level services (normally, Web Services from external providers). However, it may be difficult to establish what performance level should be required from the composed services. Too little, and the performance requirements for the composition will not be met. Too much, and the provider may charge more than desired. In addition, developers must test the external services to ensure that they can provide the required performance levels.

There is a large variety of proposals for estimating the required level of performance and measuring the actual performance of a system [1]. Measurements can be used for detecting performance degradations over time, identifying load patterns or checking the SLAs. However, the requirements set by the SLA are usually broad and cover a large amount of functionality: when violated, it might be hard to pinpoint the original cause. Ideally, we should have performance requirements for every part of the system, but that would be too expensive for all but the most trivial systems.

In our previous work [2], we presented two inference algorithms for performance annotations in workflow models. These algorithms can "fill in the blanks" for the response time and throughput requirements of every activity in the model, starting from a global annotation and some optional local annotations set by the user. Users would then write the actual performance tests manually, taking the results produced by these algorithms as a reference. However, writing these tests for every part of a reasonably-sized system could incur in a considerable cost: ideally, it should be partly automated.

In this work, we will outline how to reduce the effort involved in using the results produced by these algorithms by assisting the user in producing concrete performance tests. The models will be used to generate partial test plans and to wrap existing functional test cases as performance tests. To do so, we will weave the existing performance models with design and/or implementation models, relating the performance requirements with the appropriate software artefacts.

The rest of this work is structured as follows: after introducing the models used, we will describe our general approach for generating tests. We will then show two applications, linking the performance requirements to several kinds of software artefacts, and select several candidate technologies. Finally, we will offer several conclusions to this work, and list some future lines of work.

II. Performance models

This section will present the notation used by the performance algorithms described in [2]. We use standard UML activity diagrams, annotated with a small subset of the OMG Modelling and Analysis of Real-Time and Embedded Systems

ISBN: 978-1-61208-200-4

(MARTE) profile [3]. MARTE provides both a set of predefined performance metrics and some mechanisms to define new ones. In our case, we are using the predefined performance metrics defined in the Generic Quantitative Analysis Modelling (GQAM) subprofile. GQAM is the basic analysis subprofile in MARTE: the Schedulability Analysis Modelling (SAM) and Performance Analysis Model (PAM) subprofiles are based on it. However, SAM and PAM are outside the scope of our approach.

Figure 1 shows a simple example. Inferred annotations are highlighted in bold:

1) The activity is annotated with a ≪GaScenario≫ stereotype, in which `respT` specifies that every request is completed within 1 second, and `throughput` specifies that 1 request per second needs to be handled.
2) In addition, the activity declares a set of context parameters in the `contextParam` field of the ≪GaAnalysisContext≫ stereotype. These variables represent the time per unit of weight that must be allocated to their corresponding activity in addition to the minimum required time. Their values are computed by the time limit inference algorithm.
3) Each action in the activity is annotated with ≪GaStep≫, using in `hostDemand` an expression of the form $m + ws$, where m is the minimum time limit, w is the weight of the action for distributing the remaining time, and s is the context parameter linked to that action.
   The time limit inference algorithm adds a new constraint to `hostDemand`, indicating the exact time limit to be enforced. The throughput inference algorithm extends `throughput` with a constraint that lists how many requests per second should be handled. As these constraints have been automatically inferred, their `source` attribute is set to `calc` (calculated).
4) Outgoing edges from condition nodes also use ≪GaStep≫ but only for the `prob` attribute, which is set by the user to the estimated probability it is traversed.

III. Overall approach

The model shown in the previous section is entirely abstract: at that level of detail, it cannot be executed automatically. It will have to be implemented through other means.

After it has been implemented, it would be useful to take advantage of the original model to generate the performance test cases. However, the model lacks the required design and implementation details to produce executable artefacts. To solve this issue, several approaches could be considered:

1) The abstract model could be extended with additional information, but that would clutter it and make it harder to understand.
2) On the other hand, the implementation models could be annotated with performance requirements, but this would also pollute their original intent.
3) Finally, a separate model that links the abstract and concrete models could be used. This is commonly known as a *weaving model*. Several technologies already exist for implementing these, such as AMW [4] or Epsilon ModeLink [5].

In order to preserve the cohesiveness of the abstract performance model and the design and implementation models, we have chosen the third approach.

After establishing the required links, the next step is generating the tests themselves. To do so, a regular Model-to-Text (M2T) transformation could be used, written in a specialised language such as the Epsilon Generation Language [6]. In case it were necessary to slightly refine or validate the weaving model before, an intermediate Model-to-Model (M2M) transformation could be added. Figure 2 illustrates the models and steps involved in our overall approach.

IV. Applications

We will now show several instances of the overall approach in Figure 2, using different technologies to assist in generating performance test artefacts in different environments.

A. Reusing functional tests as performance tests

Generating executable performance test cases from scratch automatically will usually require many detailed models and complex transformations, which are expensive to produce and maintain. The initial effort required may deter potential adopters. An alternative inexpensive approach is to repurpose existing functional tests as performance tests. This is the aim of libraries such as JUnitPerf [7] or ContiPerf [8]: we will target these libraries in order to simplify the transformations involved and make the generated code more readable.

Listing 1 shows how JUnitPerf is normally used. The original *TFunctional* functional test suite is wrapped into a *TimedTest* (implemented by JUnitPerf) that checks that every test case in *TFunctional* does not take any longer than 1000 milliseconds. The wrapped test case is wrapped once again with a *LoadTest* (also implemented by JUnitPerf) that emulates 10 users running the test at the same time. In combination, the resulting test checks that each of the 10 concurrent executions of the wrapped test finishes within 1 second.

Listing 2 shows a similar fragment for ContiPerf. Instead of using Java objects, ContiPerf uses Java 6 annotations, which would be easier to generate automatically. The *@PerfTest* annotation indicates that the test will be run 100 times using 10 threads, so each thread will perform 10 invocations. *@Required* indicates that each of these invocations should finish within 1000 milliseconds at most. *@SuiteClasses* points to the JUnit 4 test suites to be reused for performance testing, and *@RunWith* tells JUnit 4 to use the ContiPerf test runner.

In both cases, the code itself is straightforward to generate. However, the generated code must integrate correctly with the existing code. If the code was not produced using a model-driven approach, there will not be a design or implementation model to link to. Instead, we will derive a model of the structure of the existing code using a model discovery tool such as Eclipse MoDisco [9]. Eclipse MoDisco can generate models from Java code such as that shown in Figure 3.

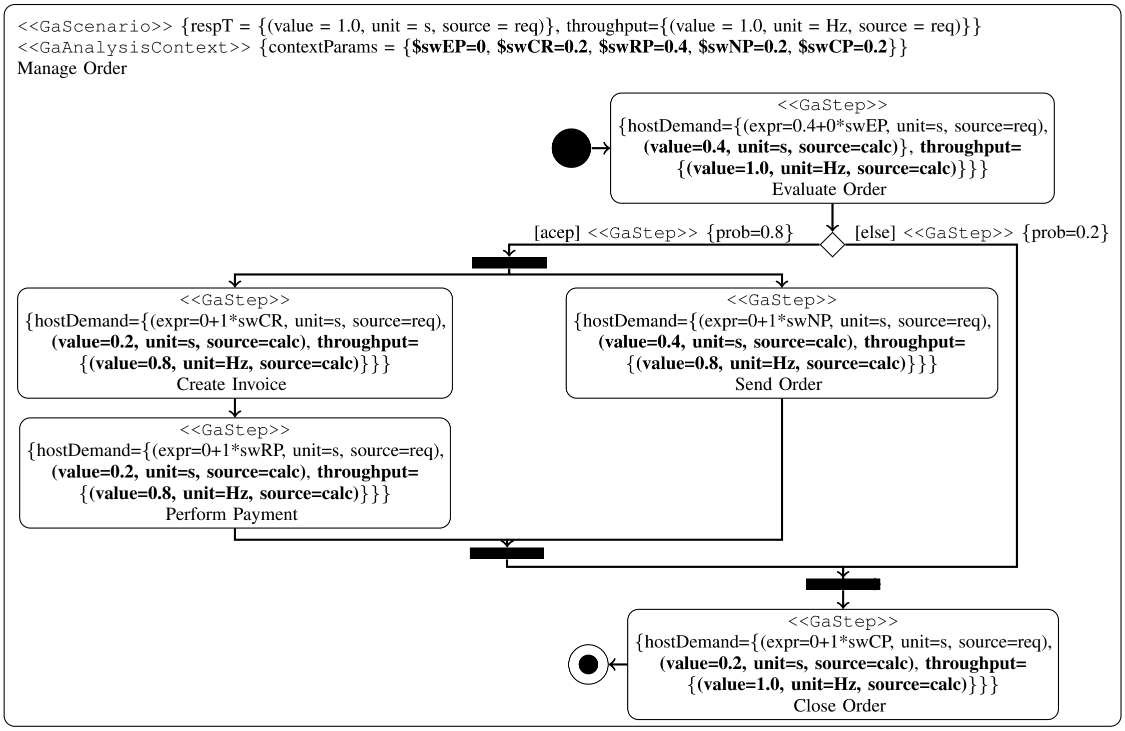

<<GaScenario>> {respT = {(value = 1.0, unit = s, source = req)}, throughput={(value = 1.0, unit = Hz, source = req)}}
<<GaAnalysisContext>> {contextParams = {**$swEP=0**, **$swCR=0.2**, **$swRP=0.4**, **$swNP=0.2**, **$swCP=0.2**}}
Manage Order
<<GaStep>>
{hostDemand={(expr=0.4+0*swEP, unit=s, source=req), **(value=0.4, unit=s, source=calc)**}, **throughput= {(value=1.0, unit=Hz, source=calc)}**}}
Evaluate Order
[acep] <<GaStep>> {prob=0.8}
[else] <<GaStep>> {prob=0.2}
<<GaStep>>
{hostDemand={(expr=0+1*swCR, unit=s, source=req), **(value=0.2, unit=s, source=calc)**, **throughput= {(value=0.8, unit=Hz, source=calc)}**}}
Create Invoice
<<GaStep>>
{hostDemand={(expr=0+1*swNP, unit=s, source=req), **(value=0.4, unit=s, source=calc)**, **throughput= {(value=0.8, unit=Hz, source=calc)}**}}
Send Order
<<GaStep>>
{hostDemand={(expr=0+1*swRP, unit=s, source=req), **(value=0.2, unit=s, source=calc)**, **throughput= {(value=0.8, unit=Hz, source=calc)}**}}
Perform Payment
<<GaStep>>
{hostDemand={(expr=0+1*swCP, unit=s, source=req), **(value=0.2, unit=s, source=calc)**, **throughput= {(value=1.0, unit=Hz, source=calc)}**}}
Close Order

Fig. 1. Simple example model annotated by the performance inference algorithms

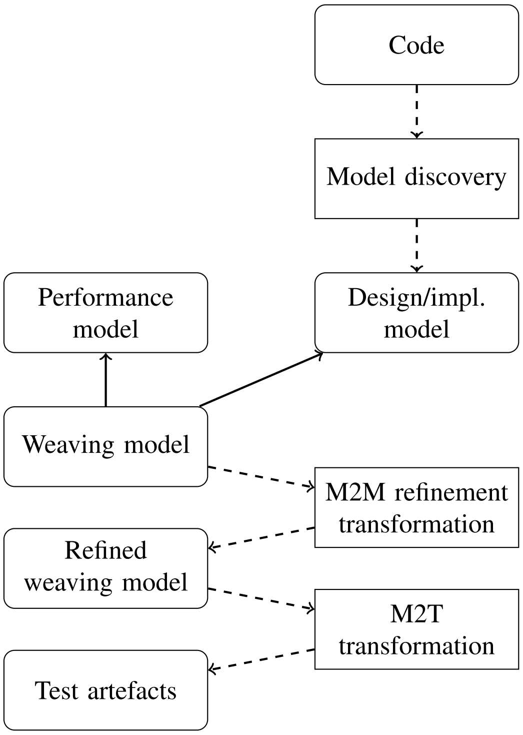


Fig. 2. Overall approach for generating performance test artefacts from abstract performance models

```
final int users = 10;
final int tlimit_ms = 1000;
Test testCase = new TFunctional();
Test test1User = new TimedTest(testCase, tlimit_ms );
Test testAllUsers = new LoadTest(test1User, users );
```

Listing 1. Java code for wrapping the *TFunctional* JUnit 3 test case using JUnitPerf

```
@RunWith(ContiPerfSuiteRunner.class)
@SuiteClasses(TFunctionalJUnit4. class )
@PerfTest( invocations = 100, threads = 10)
@Required(max=1000)
public class InferredLoadTest {}
```

Listing 2. Java code for decorating the *TFunctionalJUnit4* JUnit 4 test suite using ContiPerf

Once we have the performance and the implementation models, the next step is to link them using a new *weaving model*. Each model consists of an instance of *WeavingModel*, which contains a set of *Link*s between a ≪GaStep≫ stereotype of the MARTE performance model, and a *MethodDeclaration* of the MoDisco model. We can populate the weaving model using the standard Epsilon Modeling Framework (EMF) editors or using Epsilon ModeLink (as in Figure 4).

After linking both models with the weaving model, the last step is running a M2T transformation to produce the actual performance test artefacts. The generated code would be similar to that in Listings 1 or 2.

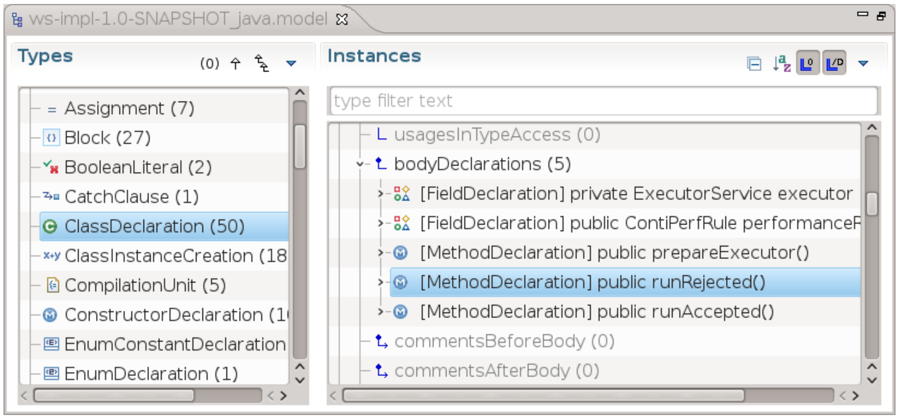

ws-impl-1.0-SNAPSHOT_java.model
Types
(0)
Assignment (7)
Block (27)
BooleanLiteral (2)
CatchClause (1)
ClassDeclaration (50)
ClassInstanceCreation (18
CompilationUnit (5)
ConstructorDeclaration (1
EnumConstantDeclaration
EnumDeclaration (1)
Instances
type filter text
usagesInTypeAccess (0)
bodyDeclarations (5)
[FieldDeclaration] private ExecutorService executor
[FieldDeclaration] public ContiPerfRule performanceF
[MethodDeclaration] public prepareExecutor()
[MethodDeclaration] public runRejected()
[MethodDeclaration] public runAccepted()
commentsBeforeBody (0)
commentsAfterBody (0)

Fig. 3. MoDisco model browser showing a model generated from an Eclipse Java project

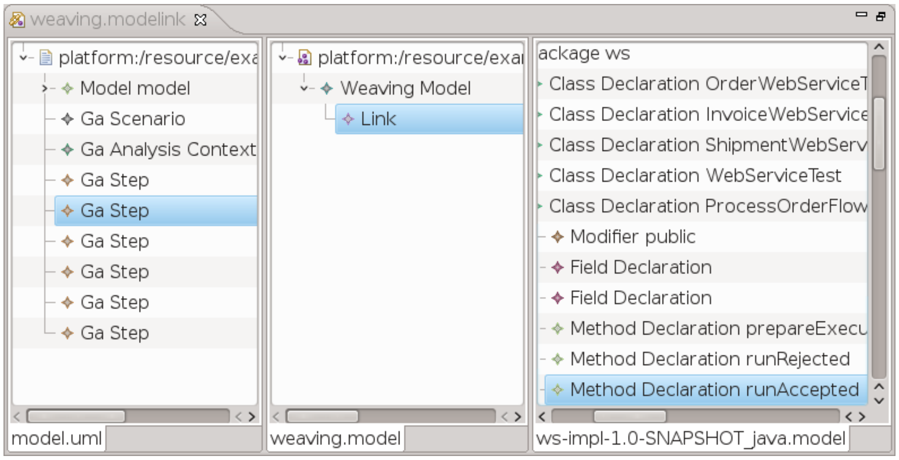

weaving.modelink
platform:/resource/exa
Model model
Ga Scenario
Ga Analysis Context
Ga Step
Ga Step
Ga Step
Ga Step
Ga Step
Ga Step
platform:/resource/exa
Weaving Model
Link
ackage ws
Class Declaration OrderWebServiceT
Class Declaration InvoiceWebService
Class Declaration ShipmentWebServ
Class Declaration WebServiceTest
Class Declaration ProcessOrderFlow
Modifier public
Field Declaration
Field Declaration
Method Declaration prepareExecu
Method Declaration runRejected
Method Declaration runAccepted
model.uml
weaving.model
ws-impl-1.0-SNAPSHOT_java.model

Fig. 4. Screenshot of the Epsilon ModeLink editor weaving the MARTE performance model and the MoDisco model

```
@WebService
public class HelloWorld {
  @WebMethod
  public String greet(@WebParam(name=”name”)
   String name)
  {
    return ”Hello␣” + name;
  }
}
```

Listing 3. Java code using JAX-WS to implement a “HelloWorld” Web Service

B. Partial test plan generation for Web Services

In the previous section, we applied our approach to existing JUnit test cases, repurposing them as performance test cases. In this section we will discuss how to generate performance test artefacts for a Web Service (WS) [10] in a language agnostic manner.

Web Services based on the WS-* technology stack are usually described using a Web Services Description Language (WSDL) [11] document. This XML-based document is an abstract and language-independent description of the available operations for the service and the messages to be exchanged between the service and its consumers. Existing Web Service frameworks such as Apache CXF [12] can generate most of the code required to implement and consume the services from the WSDL description. Users only need to implement the business logic of the services. In addition, some frameworks (CXF included) can work in reverse, generating WSDL from adequately annotated code. Listing 3 shows an example fragment of Java code that implements a simple “Hello world” Web service using standard JAX-WS [13] annotations.

Since a WSDL document is a declarative and language-independent description of the Web Service itself, we can use it as our design model. After transforming automatically the XML Schema description of the WSDL document format into a regular ECore metamodel [14], we will be able to load WSDL documents as regular Eclipse Modeling Framework models, reusing most of the technologies mentioned in Section IV-A.

The weaving model needs to relate the ≪GaStep≫ stereotypes with the operations of the services in the WSDL document. For instance, we might want to ensure that every invocation of the *evaluate* operation of the *Order* processing

```
grinder . processes=5
grinder . runs=100
grinder . processIncrement=1
grinder .  processIncrementInterval =1000
```

Listing 4. Example `.properties` file with configuration parameters for the workload

```
class TestRunner:
  def __call__( self ):
    def invoke ():
      response = HTTPRequest().POST(
        ”http :// localhost :8080/ orders”,
        ” (... ␣SOAP␣message␣...)”)
       stats = grinder .  statistics  . getForCurrentTest ()
       stats . success = (response . statusCode != 200
                          and stats .time < 150)
       test = Test (1, ”Query␣order␣by␣ID”).wrap(invoke)
       test ()
```

Listing 5. Example Jython script for The Grinder with the contents of the performance test to be run by each simulated client

service finishes within a certain time while handling a certain number of requests per second.

After weaving the WSDL-based model with the performance model, the next step is generating a test plan for a dedicated performance testing tool such as The Grinder [15]. Using a dedicated tool allows for defining tests with less cost and in a way that is independent of the implementation language of the software under test.

In the case of The Grinder, we would need to generate two different files: a `.properties` file indicating several parameters of the workload to be generated, and a Jython script with the test to be run by each simulated client. Listings 4 and 5 show simple examples for these two files. The `.properties` file in Listing 4 indicates that 5 processes should each run the test 100 times, starting with 1 process and adding one more every 1000 milliseconds. On the other hand, the test itself consists of sending an appropriate SOAP message to a specific URL and checking that the response has the OK (200) HTTP status code and that it was received within 150 milliseconds. Since these inputs are quite concise, we deem it feasible to generate an initial version of both files, letting the user add a meaningful SOAP message later.

Later iterations of this application could generate larger parts of the test plan by assisting the user in producing the messages themselves. Links in the weaving model could allow the user to specify a certain strategy for generating the messages to be sent, such as random testing, variations upon a predefined template or static analysis of the code implementing the service. The strategy could be applied in the weaving model refining step showed in Figure 2.

V. Related work

According to Woodside et al. [1], performance engineering comprises all the activities required to meet performance requirements. These activities include defining the requirements, analysing early performability models (such as layered queuing networks [16] or process algebra specifications [17]) or testing the performance of the actual system. Our previous work in [2] focused on helping the user define the requirements using MARTE-annotated [3] UML activity diagrams as notation. The present work is dedicated to helping the user create the performance test artefacts.

Our work does not deal directly with the implemented system, but rather with a simplified representation (a *model*). There is a large number of works dealing with model-based testing, i.e., “the automatable derivation of concrete test cases from abstract formal models, and their execution” [18]. Most of them (as evidenced by [18] itself) are dedicated to functional testing: we will focus on those dedicated to model-based performance testing.

Barna et al. present in [19] a hybrid approach, which uses a 2-layered queuing network (LQN) to derive an initial stress workload for a website. This workload is used to test the system and refine the original LQN model in a feedback loop that searches for the minimum load that would make the system violate one of its performance constraints. Like our work, it combines the analysis of a model with the execution of a set of test cases. However, its goal is completely different: we intend to define the appropriate quality service levels for the individual services in order to meet the desired quality service level of the entire workflow, whereas this approach would estimate the maximum workload that a workflow could handle within a certain quality service level.

Di Penta et al. show in [20] another approach with the same goal of finding workloads that induce service level agreement violations. However, they use genetic algorithms instead of a LQN model and test WSDL-based Web Services instead of a regular website.

Suzuki et al. have developed a model-based approach for generating testbeds for Web Services [21]. SLA and behaviour models are used to generate stubs for the external services used by our own service. This allows users to check that their own services can work correctly and with the expected level of performance as long as the external services meet their SLAs. However, this approach does not generate input messages for the services themselves. Still, we could use this work to check the validity of the performance constraints inferred by our algorithms in [2] in combination with the approach in Section IV-B, by replacing all services in the workflow with stubs and testing the performance of the composition.

As illustrated by the above references, there is a wealth of methods for generating performance test cases and testbeds for Web Services. However, we have been unable to find another usage of model weaving for generating performance test artefacts for multiple technologies. This is in spite of the fact that model composition using model weaving has been used regularly ever since the authors of the original ATLAS Model Weaver proposed it [4]. For instance, Vara et al. use model composition to decorate their extended use case models with additional information required for a later transformation [22].

VI. Conclusion and future work

In this work, we have described an overall approach for generating performance test artefacts from the abstract performance models produced by our inference algorithms in [2]. To generate concrete test artefacts while keeping the abstract performance models separated from any design or implementation details, we propose linking the performance model to a design or implementation model using an intermediate *weaving model*. If a design or implementation model is not available, it can be extracted from the existing code. The weaving model can be then optionally refined using a model-to-model transformation, and finally transformed into the performance test artefacts with a model-to-text transformation.

We have performed an initial study of the feasibility of the approach by studying how to apply it in two situations. The first application will reuse existing JUnit test cases as performance test cases with JUnitPerf and ContiPerf. The implementation model is extracted from the Java code implementing the test cases using the model discovery tool MoDisco [9], and the weaving model links the MARTE annotations in our performance model to the Java test methods in the MoDisco model.

The second application will generate test plans for an independent load testing framework, such as The Grinder [15]. In this case, the WSDL description of the service serves as an explicit design model, and the weaving model links the MARTE performance requirement to an operation of the service. Later revisions of this approach may use the weaving model to specify a strategy for generating the required input messages, instead of leaving it up to the user.

Our next work is to further these feasibility studies by implementing the required transformation workflows. We have implemented a considerable part of the first approach already using MoDisco and Epsilon ModeLink, and we are currently implementing the code generation step in the Epsilon Generation Language.

Acknowledgements

This work was funded by the research scholarship PU-EPIF-FPI-C 2010-065 of the University of Cádiz, the MoDSOA project (TIN2011-27242) under the National Program for Research, Development and Innovation of the Ministry of Science and Innovation (Spain) and by the PR2011-004 project under the Research Promotion Plan of the University of Cádiz.

References

[1] M. Woodside, G. Franks, and D. Petriu, “The future of software performance engineering,” in *Proc. of Future of Software Engineering 2007*, 2007, pp. 171–187.

[2] A. García-Domínguez, I. Medina-Bulo, and M. Marcos-Bárcena, “Model-driven design of performance requirements with UML and MARTE,” in *Proceedings of the 6th International Conference on Software and Data Technologies*, vol. 2. Seville, Spain: SciTePress, Jul. 2011, pp. 54–63.

[3] Object Management Group, “UML Profile for Modeling and Analysis of Real-Time and Embedded systems (MARTE) 1.0,” http://www.omg.org/spec/MARTE/1.0/, Nov. 2009, last checked on 2012-03-03.

[4] M. D. Del Fabro, J. Bézivin, and P. Valduriez, “Weaving models with the eclipse AMW plugin,” in *Proceedings of the 2006 Eclipse Modeling Symposium, Eclipse Summit Europe*, Esslingen, Germany, Oct. 2006.

[5] D. S. Kolovos, “Epsilon ModeLink,” 2010, last checked on 2012-03-03. [Online]. Available: http://eclipse.org/epsilon/doc/modelink/

[6] D. S. Kolovos, R. F. Paige, L. M. Rose, and A. García-Domínguez, “The Epsilon Book,” 2011, last checked on 2012-03-03. [Online]. Available: http://www.eclipse.org/epsilon/doc/book

[7] M. Clark, “JUnitPerf,” Oct. 2009, last checked on 2012-03-03. [Online]. Available: http://clarkware.com/software/JUnitPerf.html

[8] V. Bergmann, “ContiPerf 2,” Sep. 2011, last checked on 2012-03-03. [Online]. Available: http://databene.org/contiperf.html

[9] H. Bruneliere, J. Cabot, F. Jouault, and F. Madiot, “MoDisco: a generic and extensible framework for model driven reverse engineering,” in *Proceedings of the IEEE/ACM International Conference on Automated Software Engineering*, Antwerp, Belgium, Sep. 2010, pp. 173–174.

[10] H. Haas and A. Brown, “Web services glossary,” World Wide Web Consortium, W3C Working Group Note, Feb. 2004, last checked on 2012-03-03. [Online]. Available: http://www.w3.org/TR/ws-gloss/

[11] World Wide Web Consortium, “WSDL 2.0 part 1: Core Language,” http://www.w3.org/TR/wsdl20, Jun. 2007, last checked on 2012-03-03.

[12] Apache Software Foundation, “Apache CXF,” Nov. 2011, last checked on 2012-01-15. [Online]. Available: https://cxf.apache.org/

[13] Java.net, “JAX-WS reference implementation,” Nov. 2011, last checked on 2012-03-03. [Online]. Available: http://jax-ws.java.net/

[14] D. Steinberg, F. Budinsky, M. Paternostro, and E. Merks, *EMF: Eclipse Modeling Framework*, 2nd ed., ser. Eclipse Series. Addison-Wesley Professional, Dec. 2008.

[15] P. Aston and C. Fizgerald, “The Grinder, a Java Load Testing Framework,” 2012, last checked on 2012-03-03. [Online]. Available: http://grinder.sourceforge.net/

[16] D. C. Petriu and H. Shen, “Applying the UML Performance Profile: Graph Grammar-based Derivation of LQN Models from UML Specifications,” in *Proc. of the 12th Int. Conference on Computer Performance Evaluation: Modelling Techniques and Tools (TOOLS 2002)*, ser. Lecture Notes in Computer Science. London, UK: Springer Berlin, 2002, vol. 2324, pp. 159—177.

[17] M. Tribastone and S. Gilmore, “Automatic extraction of PEPA performance models from UML activity diagrams annotated with the MARTE profile,” in *Proc. of the 7th Int. Workshop on Software and Performance*. Princeton, NJ, USA: ACM, 2008, pp. 67–78, last checked on 2012-03-03. [Online]. Available: http://portal.acm.org/citation.cfm?id=1383569

[18] M. Utting, A. Pretschner, and B. Legeard, “A taxonomy of model-based testing,” Working Paper 04/2006, Apr. 2006, last checked on 2012-03-03. [Online]. Available: http://researchcommons.waikato.ac.nz/handle/10289/81

[19] C. Barna, M. Litoiu, and H. Ghanbari, “Model-based performance testing (NIER track),” in *Proceedings of the 33rd International Conference on Software Engineering*, ser. ICSE ’11. New York, NY, USA: ACM, 2011, pp. 872–875.

[20] M. Di Penta, G. Canfora, G. Esposito, V. Mazza, and M. Bruno, “Search-based testing of service level agreements,” in *Proceedings of Genetic and Evolutionary Computation Conference*, H. Lipson, Ed. London, United Kingdom: ACM, Jul. 2007, pp. 1090–1097.

[21] K. Suzuki, T. Higashino, A. Ulrich, T. Hasegawa, A. Bertolino, G. De Angelis, L. Frantzen, and A. Polini, “Model-based generation of testbeds for web services,” in *Testing of Software and Communicating Systems*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, vol. 5047, pp. 266–282.

[22] J. M. Vara, M. V. De Castro, M. Didonet Del Fabro, and E. Marcos, “Using weaving models to automate model-driven web engineering proposals,” *International Journal of Computer Applications in Technology*, vol. 39, no. 4, pp. 245–252, 2010.

Virtual Web and Experimental Vibration Laboratory Coordination as an Educational Tool for Mechanics Teaching in Engineering

Martínez Valle, José Miguel, Balcaza Bautista, Teresa y Martínez Jiménez, José Miguel
Mechanics Department
UCO, Polytechnic School
Córdoba, Spain
e-mail: jmvalle@uco.es

Martínez Jiménez, Pilar, Pedrós Pérez, Gerardo
Applied Physics Department
UCO, Polytechnic School
Córdoba, Spain
e-mail: fa1majip@uco.es

***Abstract*—The purpose of this paper is to evaluate the effectiveness of using virtual laboratory as a complementary tool for mechanic teaching. The use of Virtual Oscilloscope Web and Virtual Damped Oscillations Laboratories, together with the study of the vibration of beams Experiment in a real Laboratory in the university teaching system, are explained. The computer programmes can be used from the online and download areas of the Applied Physics and Mechanics Departments internet server, in order to be executed via web. These computer applications permit one to simulate practice behavior, and the users can work step by step in the same way as they do in the real laboratory, obtaining the corresponding calculations and plots. The Virtual Laboratory Web has been developed by our research team and the main objective is to familiarize the student with the oscilloscope and the Damped Oscillations experiment for its later handling in a real laboratory. The aim of these tools is to help students to learn, to study, and to investigate on their own. Furthermore, students can prepare their experiment lessons before going into the laboratory and revise them whenever, and as many times as they want to. Consequently, in using the computer as a complementary educational tool, the quality of university teaching is greatly improved.**

Keywords–Education; Technology; Simulation Software; Multimedia; Oscillations; Virtual Laboratory.

I. INTRODUCTION

In the past twenty years, there has been widespread experimentation with computer-assisted didactic models in the training of highly qualified experts such as aircraft pilots, astronauts and engineers specialized in controls. The same has happened in teaching centers using models, which allow the computer to emulate the working of different laboratory equipment and to instruct students, how to use it.

The carrying out of these experiments in the Mechanics teaching laboratory is very costly and requires students to devote considerable time to the study of each of the experiment facilities and to the handling of the equipment. In the Laboratory of the Mechanics Department at Córdoba University, we have two real vibration bench practices that we use to study the natural frequencies of simple beams (Isostatic beams as cantilevers, articulated beams, etc.) with or without damping vibrations.

However, due to severe overcrowding in classrooms and not enough necessary time, we have been developing and using a virtual lab that shows the real practices to the student.

This paper presents a new learning experience of the coordination between the virtual laboratories and the real experiment in the beams vibration practice.

Firstly, we describe two virtual laboratories developed by our research team: the virtual web oscilloscope and the virtual Damped Oscillations laboratory.

Furthermore, we present their application to the practice in the Mechanics Laboratory and we analyze the results obtained by the students in the practice classes.

II. STATE OF THE ART

Currently, a lot programmes for the simulation of certain important Physics experiments have been created [1], whose worth lies in contributing to the students' capacity to perform, in a simplified manner, those mental actions which are similar to those that they might have to carry out in a traditional laboratory, with the aid of these models.

From a general point of view, numerous applications exist whose main objective is education and the transmission of knowledge [2]. The use of the computer for such an aim accelerates the learning process of the concepts dealt with, since the interaction with the user allows them to be assimilated in a more rapid and easier way [3].

In the field of scientific and technological education, the computer can be used as a reflective tool, where the student is a protagonist of his own learning process [4] [5].

From this perspective, our line of work has focused on the development and evaluation of applications that include different modules: diagnoses of knowledge and previous ideas, resolution of problems, numerical simulations, virtual laboratories, interactive tutorials, etc. From an educational point of view, the principal didactic usefulness of the tool presented is that simulations of the phenomena and virtual experiments offer a certain degree of realism so that the student can modify the independent variables or the initial conditions and can analyze the changes that take place in the systems [6].

The development in the Departments of Applied Physics and Mechanics at the Polytechnic School of Córdoba University (Spain) of computer applications for didactic purposes began during the academic year of 1988-1989. From the 1991-1992 academic year onwards, the imparting of practices simulated together with the experimental practices has been generalized. This methodology is applied in the learning of Physics and Mechanics in the first years of Industrial Engineering and Computer Engineering [7].

For all these reasons, our group has been involved in the development, application and evaluation of virtual laboratories and their utilization, with a series of encouraging results that have been described in other works [8].

III. Objectives

The purpose of this study is to evaluate the effectiveness of using virtual laboratory as a complementary tool for mechanic teaching. For that, an objective is to coordinate virtual web and experimental laboratories.

Also, another objective is to familiarize the student with the oscilloscope and the Damped Oscillations experiment for its later handling in a real laboratory.

These tools help students to learn, to study, and to investigate on their own.

The computer applications include Interactive tutorials, and Questionnaires, which give students the possibility of evaluating themselves.

IV. Teaching Experiment Description

Before doing the practices, the student downloads the programme guide from the learning web of the Mechanics Department, in which a brief introduction is given on the theoretical fundamentals, recommending the student to consult the virtual laboratories [9] [10], the tutorials [11], the apparatus, and the operating mode of the experiment.

On the day that the experiment takes place; the student has to take a brief summary to the laboratory to be handed in the moment the experiment begins. If possible, the student is recommended to take a portable computer to the laboratory so that there is one per 2 students (one for each work post). The teachers also provide them with two computers.

The students carry out the experiment processes by consulting the operating mode in the virtual laboratories and checking the results obtained experimentally with those that they would obtain by simulation. Finally, students have to hand in a final report of the practical, in which they include the data treatment, results obtained and conclusions.

V. Software technical details on the tool/system

The damped oscillation software has been created using Visual Basic 5.0 language, with the aim of making the user feel that he/she understands and has a good command of it since we have tried to make it very intuitive and easy to use. The Oscilloscope programme has been created with Action Script and it permits user-software maximum interaction via web.

Both programmes are in research team VLC web server [10] [11].

VI. Software PROPERTIES

The interface of the computer applications must be the simplest, most intuitive and most attractive possible, so that it allows the user to interact with the machine and to obtain an ideal execution of the presentations that the application developed offers [12]. The characteristics of our system are described as follows:

*It facilitates the user's browsing for the different parameters included in the application; presenting them in a form arranged in such a way as to avoid confusion.

*The elements of the interface are accommodated so that their position on the screen facilitates the transition between the thought of the user and the action to be carried out.

*The application is intuitive and attractive.

*There is an exhaustive control of erroneous information so that mistaken results cannot be returned to the users without providing informative messages about the mistake committed.

VII. Objectives of The Implemented Softwares

The primary target of this work has been to remedy the deficiencies of Technical Studies students at the time of acquiring knowledge. Concretely, in doing practice work in laboratories, especially in that corresponding to experimentation with beam vibrations , for which it is necessary to know how to use with ease an oscilloscope and the experimental handling of damped oscillation equipment.

By means of the virtual oscilloscope, which is included within a more global project of virtual laboratories encompassing different Physics and Engineering problems, , the student can become familiar with the handling of an oscilloscope (Figure 1). Other objectives achieved are:

•Objective 1: Multimedia has been developed for a system that includes sample information necessary for students in the Vibrations field.

• Objective 2: It allows the user to become familiar with some of the devices used in the oscilloscope.

• Objective 3: “Multiplatform” works in surroundings; in addition, the application can be used under any platform, or Linux, Windows, etc.

• Objective 4: To reproduce schemes of the real systems so that the students can visualize the problems created.

VIII. Description Of The Virtual Laboratory

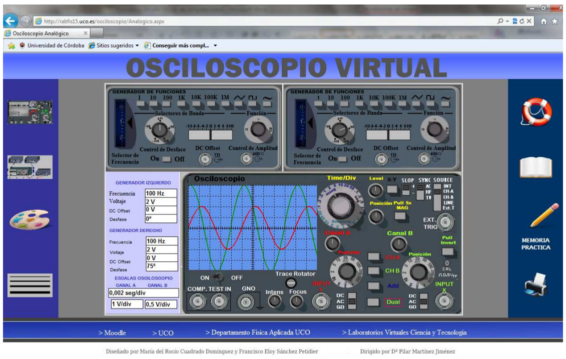
A. *Oscilloscope Virtual Laboratory*

The Software [10] is structured in three different sections: Tutorial, Simulation, and Help.

General Interface. The interface of a computer system must be as simple, intuitive and attractive as possible, so as to allow the user to interact with the machine.

Theoretical Tutorial Module. The tutorial is clear and concise, using pictures and diagrams, the basic concepts of the subject matter, the study's scope and its application to the vibrations of simple beams.

ISBN: 978-1-61208-200-4


http://rabfis15.uco.es/osciloscopio/Analogico.aspx

GENERADOR IZQUIERDO

Frecuencia	100 Hz
Voltaje	2 V
DC Offset	0 V
Desfase	0°

GENERADOR DERECHO

Frecuencia	100 Hz
Voltaje	2 V
DC Offset	0 V
Desfase	75°

ESCALAS OSCILOSCOPIO

CANAL A	CANAL B
0,002 seg/div	
1 V/div	0,5 V/div

Diseñado por María del Rocío Cuadrado Domínguez y Francisco Eloy Sánchez Petidier

Dirigido por Dª Pilar Martínez Jiménez

Figure 1. Virtual Oscilloscope

Simulation. The main objective of this lab is to learn to use the analog oscilloscope to display and measure periodic signals in time. For this, as shown in the image of the virtual oscilloscope, we have different switches and controls that allow us to modulate the frequency, the amplitude of the signal, and so on.

B. *Damping Oscillation Simulation Lab*

The study of the vibrations in beams is implemented by an experimental bench with springs.

The practice laboratory that we have implemented in the computer basically consists of a frame on which springs with different elastic constants can be suspended and loads also added. A piston is placed in a vessel which is filled with liquids of different viscosities and a recorder equipped with a pen traces the different types of motion: quasi-free, under damped, critically damped and over damped oscillations. In actual fact, we have tried to simulate the real vibration bench practice shown in Figure 2.

The essence of this application consists of the design and creation of interactive software incorporating the most important experiments that can be done by students with a free and damped vibration bench. This comprises four parts, which can be accessed from the main menu: Tutorial, Simulation, Introduction and on-line Help.

In the simulation module, a study can be made of the motion of a load suspended vertically from a spring in terms of the following parameters: viscosity, system load and elastic constant of the springs. This module is a virtual representation of the instruments necessary for the student to be able to have it recognized and learn how to use the real vibration bench when he/she has completed this practical.

In order to create the multimedia system, a video camera film was made. It displayed both the components of the practical (loads to attach, springs, oils, etc.) and the experimental process itself. The application generates graphic and numerical results. The button "Print" enables the student to print the register of the corresponding motion. The button "Following" makes it possible to go on to any new motion resulting from having modified the conditions of the system. (Figure 3)

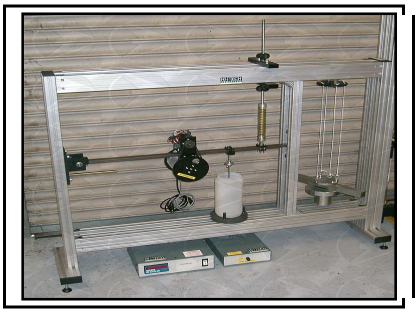
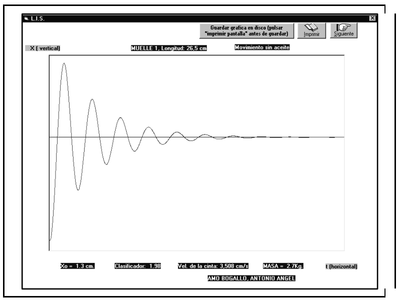


Figure 2. The real vibration bench practice


MUELLE 1, Longitud: 26,5 cm

Movimiento sin aceite

Xo = 1.3 cm.

Clasificador: 1.98

Vel. de la cinta: 3,508 cm/s

MASA = 2.7Kg.

AMO BOGALLO, ANTONIO ANGEL

L.I.S.
Guardar grafica en disco (pulsar "imprimir pantalla" antes de guardar)
Imprimir
Siguiente
X (vertical)
t (horizontal)

Figure 3. Screen for damped motion generation

IX. Educational Applications

We have used these virtual laboratories as complementary tools to traditional teaching methods, with a view to obtaining more personalized teaching to counteract the present overcrowding of university classrooms.

With the aim of checking the degree of influence of the coordination of virtual and experimental laboratories on the learning process, the results of the didactic experiments carried out during the last two academic years were compared.

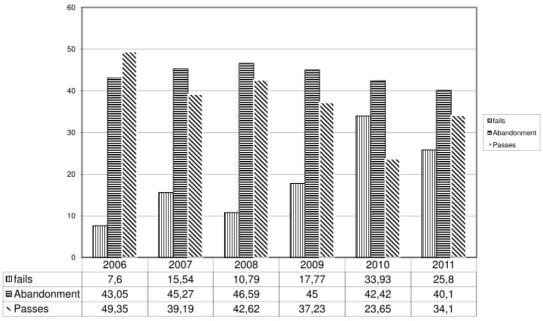
Of the 160 students enrolled in each academic year, approximately 100 signed on for the laboratories web. This was an indispensable condition for carrying them out.

Of the 100 students who began the course, approximately 90% did the real and virtual experiments, subsequently handing in their reports. This is also an indispensable condition for passing the subject "Mechanics in Engineering".

The assessment of the programmes used was made by evaluating the individual reports of the students at the end of the experiment and some complementary questionnaires on the topics tackled.

This evaluation was analyzed by classifying the results obtained by the students and establishing three categories of knowledge, i.e., abandonment of studies, fails and passes.

The results obtained show (Table 1) that, in the past two years, the number of students abandoning the experiment and virtual laboratories, and, therefore, the subject has dropped. What is more, in the same period during which the virtual Laboratories were increased and coordinated (2010 and 2011), the percentage of students failing diminished, and the amount of students with improved results in practical works rose. Thus, it can be concluded that the putting into practice and implementation of virtual laboratories in coordination with experimental ones triggers a great improvement in teaching.



	2006	2007	2008	2009	2010	2011
▥ fails	7,6	15,54	10,79	17,77	33,93	25,8
▤ Abandonment	43,05	45,27	46,59	45	42,42	40,1
▧ Passes	49,35	39,19	42,62	37,23	23,65	34,1

Table 1. Comparison of overall results for subject of Mechanics

X. Conclusions

This paper has presented the computer applications designed for didactic purposes in university teaching of a scientific and technological nature, which are currently being used and evaluated – in real education contexts – with first-year students of industrial engineering studies at Cordoba University (Spain).

The programmes are equipped with complete tutorials, presented with hypertext and images in order to help students to understand the concepts. These tutorials include animations, sound and videos, which increase their appeal to students, who can browse through the different parts of the Tutorial by means of hyperlinks and buttons connecting different parts of the system.

The simulation and virtual laboratories of real experiments are undoubtedly the most important items of these applications. Using them, students can actively interact: by incorporating data into the independent variables, in the observation of experiments, in the analysis of the results obtained, and in other aspects related to the solution of any problems encountered.

These computer applications are compact, intuitive, and easy-to-use tools, which combine, in a single application, the main elements involved in the education process: theoretical contents, practical activities (solution of problems, simulations, and virtual experiments), and the evaluation of previous or acquired knowledge.

The computer applications have been used as a learning tool incorporating it into the tasks of the course.

The evaluation of the results of student grades has shown that the use of computer simulation tools encourages students' interest in the subject and improves their marks helping to raise their level of knowledge of theoretical concepts and practical teaching techniques and problem solving.

Finally, in future projects we hope to develop 3D Virtual Labs as a relevant solution to immerse the students in a realistic experimental environment.

References

[1] H. Li and P. E. D. Love "Use of Visual Simulation in construction technology education" , Computer Applications in Engineering Education, vol. 6, no. 4, Dec. 1998, pp. 217–222, doi: 10.1002/science1099-0542.

[2] V.M. Becerra, "Solving optimal control problems with state constraints using nonlinear programming and simulation tools", IEEE Transactions on education, vol. 47 no. 3, Aug. 2004, pp. 377-384, doi: 10.1109/TE.2004.825925.

[3] E. Ras, R. Carbon, B. Decker, and J. Rech "Experience Management Wikis for Reflective Practice in Software Capstone Projects" , IEEE Transactions on education, vol. 50, no. 4, Nov.2007, pp. 312-320, doi: 10.1109/TE.2007.904580.

[4] WJ. Lee, JC. Gu, and RJ. Li, "A physical laboratory for protective relay education" IEEE Transactions on education vol. 45, no. 2, May 2002, pp. 182-186, doi: 10.1109/TE.2002.1013885. .

[5] M. Stefanovic, M. Matijevic, and V. Cvijetkovic, "Web-Based Laboratories for Distance Learning", International Journal of Engineering Education , vol. 25, no. 5, Dic. 2009, pp. 1005-1012.

[6] P. Martínez-Jiménez, M. Varo, MC. García, G. Pedrós Pérez, JM. Martínez-Jiménez, R. Posadillo, R., and EP. Varo-Martínez, "Virtual web sound laboratories as an educational tool in physics teaching in engineering", Computer Applications in Engineering Education, vol. 9, no. 4, Dic. 2011, pp. 759-769, doi: 10.1002/science1099-0542.

[7] J. Leon Alvarez, P. Martínez-Jiménez, and A. Pontes-Pedrajas, "Simulación mediante ordenador de movimientos bidimensionales en medios resistivos" Enseñanza de las Ciencias, Vol. 1, Dic. 1994, no. 12 pp. 30-38.

[8] MD. Redel-Macias; P. Martínez-Jimenez, and A. Cubero-Atienza, "E-learning applied for training on safety and hygiene in electronics engineers degree", Proc. 3 The International Conference on Computer Supported Education , (CSEDU 2011) , may 2011, pp. 258-263 ISBN 978-989-8425-49-2.

[9] http://www3.uco.es/m1112, March, 2012

[10] http://rabfis15.uco.es/osciloscopio/ March, 2012

[11] http://rabfis15.uco.es/lvct/index.php?q=node/22 March, 2012

[12] N.M. Avouris, N. Tselios, and E.C. Tatakis, "Development and evaluation of a computer-based laboratory teaching tool"; Computer Applications in Engineering Education, vol.9, no.1, April 2001, pp. 8-19, doi: 10.1002/science1099-0542.

E-Learning with Hands-On Labs in Higher European Education

Fritz Laux
Fakultät Informatik
Reutlingen University
D-72762 Reutlingen, Germany
fritz.laux@reutlingen-university.de

Martti Laiho
Dpt. of Business Information Technology
Haaga-Helia University of Applied Sciences
FI-00520 Helsinki, Finland
martti.laiho@haaga-helia.fi

Thomas Connolly
School of Computing
University of the West of Scotland
Paisley PA1 2BE, UK
thomas.connolly@uws.ac.uk

***Abstract*—”Learning by doing” has incontestably the highest enduring and motivating effects in learning. It challenges the exploratory aptitude and curiosity of a person. In higher education in information technology, exploratory learning is hindered by technical situations that are not easy to reproduce and to verify. Technical skills are, however, mandatory for employees in this area. On the other side, theoretical concepts are often compromised by commercial implementations. The challenge is to contrast and reconcile theory with practise. In two European Union funded projects we designed, implemented, and evaluated a unique e-learning approach, which realises a modularised teaching concept that provides easily reproducible virtual hands-on labs. The novelty of the approach is to use software products of industrial relevance to compare with theory and to contrast different implementations. Pilot applications in several European countries demonstrated that the participants gained highly sustainable and profound understanding about the learning objects.**

***Keywords*-learning by doing, virtual laboratory, hands-on lab, e-learning concept.**

I. Introduction

Aristotle already promoted ”learning by doing” in his eminent work on ethics, the Nicomachean Ethics [1]. The concept became known in pedagogy through the work of Comenius [2]. From the perspective of developmental biology learning by doing is known even from animals [3] and experimenting (the systematic learning by doing) is fundamental in the development of the *homo sapiens* [4][5].

Effective knowledge transfer at Higher Education (HE) institutions and Vocational Educational Training (VET) should be tailored to the needs of its clients. Employees are highly motivated to acquire new skills but are often hindered to follow a scheduled training programme. Students face a denser curriculum due to the Bologna process with a high degree of optional courses whose schedules and prerequisites are not aligned. Therefore it is essential to provide self study courses with small module sizes to enable the participants to learn in their spare time at an individual pace. In addition, in financially difficult times, knowledge transfer should be highly scalable in terms of costs. E-Learning offers this capability but has the difficulty to keep motivation high.

As consequence, e-learning has to solve a multidimensional problem: The learning content needs to be chunked into ”digestible” portions while keeping the necessary context. Technological reality has to match with the theoretical underpinning. Technological aspects in Information and Communication Technology (ICT) are of particular importance to empower students and employees for a competitive labor market. This will stimulate the secondary motivation of the learners.

In our case we focus on one of the most important areas in ICT competency for information management professionals: *database systems*. Databases are now the underlying framework of information systems that have fundamentally changed the way organizations and individuals structure and handle information.

One crucial competence within the database domain is how to structure efficiently a database and how to correctly process the data. For example, in the case of a banking application the database has to process correctly and reliably the financial transaction under any circumstances.

This requires a sound understanding of the theory and practical skills of software products at the same time. Such a highly specialized knowledge cannot be only theoretically taught neither could it be trained only by examples like a cookbook. This is the scenario for our e-learning based concept with hands-on labs.

A. Structure of the Paper

With the following overview on related work in cognitive science the context for our learning theory will be settled. In Section II we point out the pedagogical requirements, the modularization constraints dictated by the learning object and the stress field between industry demands and long term knowledge for the students. This clarification is used in Section III as criterion for developing a unique reference model for the example learning object *database systems*.

Section IV describes the supporting technology, in particular, the environment for the hands-on labs. Our findings during pilot runs of the learning modules are presented and discussed in Section V. We end the paper with a conclusion.

B. Related Work

E-learning is a promising research subject and there is an abundance of publications on the foundation of on-line learning (e.g., [6][7][8][9][10]) as well as on problems. For instance, the decreasing motivation was described by Prenzel [11] and Paechter et al. [12]. It is also confirmed by our own experience with e-learning.

According to the constructionism [13] the learner generates knowledge by individual experience (radical constructivism [14]) or by social interaction within a cultural context (social constructivism [7]). As consequence, knowledge should be acquired by the learner in authentic situations that keep motivation high [15]. Connolly and Begg [16] report similar experiences and recommend teaching database analysis and design in a problem based environment.

Communication with fellow students and team work supports motivation, too [9]. This makes a communication and collaboration tool an indispensable ingredient of an e-learning system.

Multimedia support through E-learning systems is a an enabler for flexible and scalable HE and VET, but is no guaranty for a successful on-line course. Critical voices raised the issue of superficial and routine knowledge that may easily be transferred. This knowledge refers to the cognitive domains one (knowledge) and two (comprehension) of Bloom's taxonomy [8]. Bloom's knowledge taxonomy was chosen because it fits well into the evaluation of skills related learning. But, profound insights (analysis, synthesis, and evaluation in Bloom's categories) are difficult to convey with a computer based learning environment as the study conducted by Spannagel [17] reveals.

It seems difficult to ensure that theory and the necessary abstraction are drawn from an example. There are concepts that try to overcome these problems with the use of multimedia technology [10].

Blended learning, for example, tries to combine classroom learning with e-learning ([6], chap. 10 and 29). Classroom teaching can provide for theory and the e-learning session practise the knowledge in form of exercises or experiments. We apply this technique for our virtual laboratory workshops described in Subsection III-C. This hybrid learning does not ensure sustainable and deep understanding, but, a well thought concept may help to convey deep insights as Astleitner and Wiesner [9] point out.

Our concept aims further: It contrasts and reconciles theory with the reality of commercial software products. This is important because software professionals and experts need the competence to verify the real behavior of a database system for instance and compare it with the theory. As consequence real products are necessary as training tools and for assessment. No learning concept, so far, has tried to deal with the peculiarities of commercial software products.

II. Problem Description and Contribution

The goal is to provide a highly modularized e-learning environment for the specific theoretical and practical needs of HE and VET in the domain of ICT. For the proof of concept we have chosen the material produced during two EU funded projects: DBTech Pro (funded by the Leonardo da Vinci programme) and its successor DBTech EXT (funded by the EU Lifelong Learning Programme). The content focus was on in depth knowledge with hands-on labs for database design, transaction processing, and data mining. More information about both projects may be found at http://www.dbtechnet.org.

From the pedagogical view we identified the following requirements:

- self controlled learning
- authentic problem oriented learning
- most effective, cooperative learning
- self assessment
- feedback and evaluation

Self controlled learning is important because of the above mentioned time constraints and with regard to different precognitions of the learners. For high motivation it is necessary to pose authentic, real world problems to solve [16]. This requires state-of-the-art software used in industry.

Cooperative learning has two positive effects, one for the learner and one for the teacher: Communication among the students and working in groups keep motivation high and yield better learning results. From the teacher's view the communication provides feedback on the effectiveness of the teaching and exercise material. In addition, communication among students reduces teacher intervention.

Memorized knowledge may be assessed easily through multiple choice tests but constructive tasks and creative work are a challenge to assess in a automated way.

From the skills and competences demanded by employers the following requirements need to be taken into account:

- ability to solve real world tasks (problem solving)
- knowledge about state-of-the-art technology
- social skills, so called soft skills

Employees and students have increasing interest in learning skills that give a fast and easy to see return on their learning investment in form of directly applicable knowledge at their working place. This validates the first two qualification requirements. Problem solving and social skills are indispensable for highly demanding ICT jobs [18].

In addition to the above requirement, the teaching units (modules) need to comply with the taxonomy of that domain, which defines how to slice the content along the aspects:

- competence level
- subject area
- technology

Cutting the content along the competence level provides different degrees of detail in line with target competencies and work profile. Students of HE institutions prefer a different learning concept than in VET courses. The latter have a tighter time schedule with less time for reflection of theoretical issues than HE students.

So, apart from the challenging content we tried to address all of the above requirements by slicing the learning content so that it can be combined and composed in multiple ways.

A. *Contribution*

The contribution of this paper consists of an integrated learning concept for e-learning addressing the needs and constraints of HE and VET. For each learning unit the most appropriate learning concept was applied. Furthermore, the framework solves the problem of content modularization. Exemplary e-learning material that was used in multiple pilot runs proofed the usefulness and superior knowledge sustainability compared to traditional university teaching. The main advantage lies in the practical skills acquired using real DBMS products in the hands-on labs. The necessary lab environments are easy reproducible and provide full control of license restrictions.

III. The Reference Model

The reference model applies different learning concepts reflecting the different aspects and challenges presented in the previous Section. The interrelation of these requirements make it difficult to optimise the learning concept. For better understanding we treat the dimensions content, lab environment, and project work separately and discuss the global optimisation in Subsection III-E at the end of this section.

A. *Knowledge Taxonomy*

It is common to define a syllabus for the learning content. Structuring the syllabus results in a knowledge taxonomy of the teaching domain. From this structure we are able to deduct pre-requisites, identify learning elements, and designate learning outcomes. Structuring the teaching domain along the knowledge levels defined by Bloom [8] helped us to modularize the content according to knowledge depth and to provide teaching units for different target groups. As an example, Figure 1 shows an cutout of the DBTech database taxonomy [19] showing the comprehension levels. From this layering we were able to deduct pre-requisites for every learning unit. For instance the unit *data modeling* (see Silberschatz et al. [20]) requires knowledge about the *relational, hierarchical*, and *network model*.

B. *Virtual Laboratory*

The most important component of our e-learning model is the ”learning by doing”. The psychomotoric learning keeps motivation high and supports a high degree of practical skills needed by companies. Moreover, the endurance of knowledge is much better and profound than without hands-on labs. Small, practical exercises and experimenting prepares the way for problem based learning.

In the case of ICT we have to deal with sophisticated, interdependent software systems like database management systems, application servers, data warehousing, OLAP systems, or business intelligence suites. A student would need excessive time to install and set up the lab environment. This is unfeasible, considering only the risk that the system might be (unconsciously) misconfigured.

An other obstacle could be inhomogeneous hardware that might impede the installation of a certain product. The only technical solution that works without problems is the virtualization technology. It provides a lab environment independent of the physical computer, which can be copied across the Internet to computers of the learners. Even if a student accidentally damages the virtual system he can reset it to its original state. He is also able to save his results in a snapshot and continue later or at a different computer. There exist virtual image capturing and playing software that is freely available.

C. *Virtual Laboratory Workshops*

The technological complexity of the Virtual Laboratory makes it necessary to provide detailed, step-by-step tutorials for experimenting. In order to make the learning more effective, we decided to use blended learning techniques and gather students for live workshops using the virtual laboratory. One trainer for about 10 students was sufficient to answer questions or to provide help with the virtual lab environment.

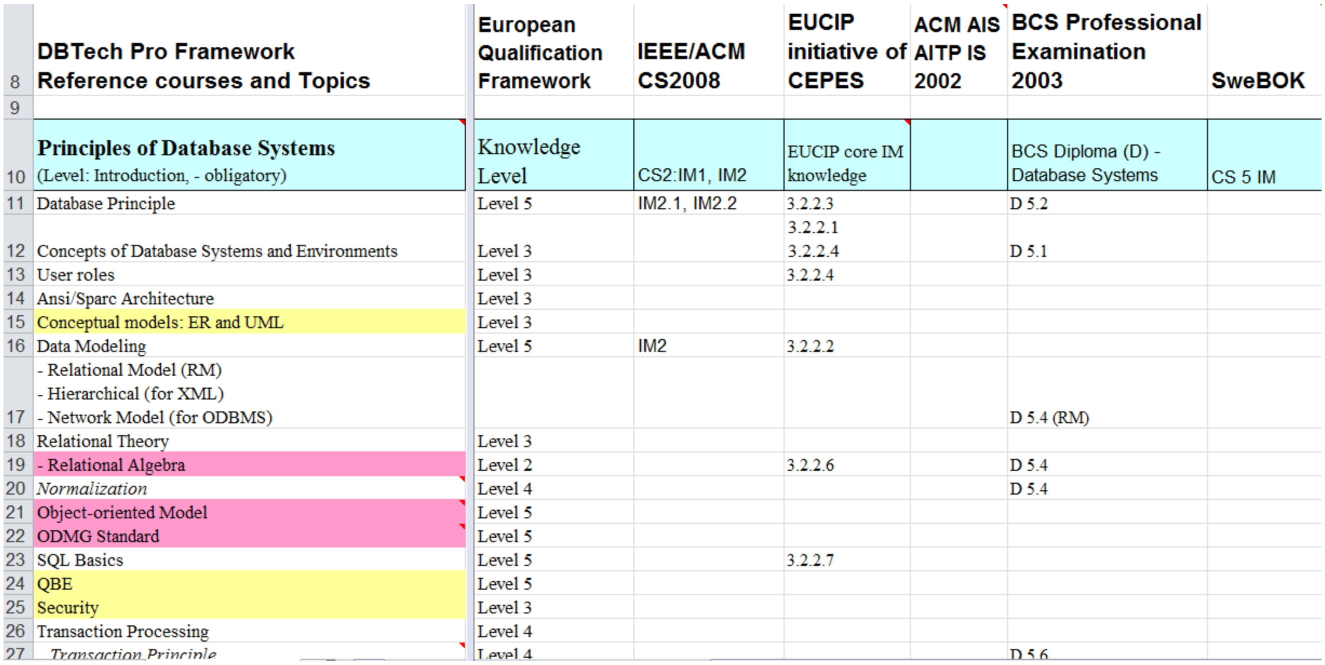
Between workshop sessions and for remote participants Skype telephone and remote assistance via web conferencing tools have been available. This allowed interactive help directly with the laboratory environment.

The students had to submit their deliverables electronically via the e-learning platform for grading. The e-learning system was also heavily used as a discussion board and for feedback from students. The feedback was used for improvements.

D. *Project Work*

While teaching theory in a didactic way and practising or verifying the transferred knowledge in hands-on labs there is no guaranty that the students really acquire a problem solving competence. It is necessary to combine different knowledge pieces, then abstract and apply them as a whole. This systemic knowledge gap can be easily seen when students know about the ACID properties [21] of a transaction, but cannot relate a real world problem like the concurrent on-line reservation of flights with the concurrency issue. In the lab with real products it is possible to test the behavior of the used software also in case of concurrent clients.

Moreover, students might be skilled in technological aspects of application servers but do not realize the danger



8	DBTech Pro Framework Reference courses and Topics	European Qualification Framework	IEEE/ACM CS2008	EUCIP initiative of CEPES	ACM AIS AITP IS 2002	BCS Professional Examination 2003	SweBOK
9							
10	**Principles of Database Systems** (Level: Introduction, - obligatory)	Knowledge Level	CS2:IM1, IM2	EUCIP core IM knowledge		BCS Diploma (D) - Database Systems	CS 5 IM
11	Database Principle	Level 5	IM2.1, IM2.2	3.2.2.3		D 5.2	
12	Concepts of Database Systems and Environments	Level 3		3.2.2.1 3.2.2.4		D 5.1	
13	User roles	Level 3		3.2.2.4			
14	Ansi/Sparc Architecture	Level 3					
15	Conceptual models: ER and UML	Level 3					
16	Data Modeling	Level 5	IM2	3.2.2.2			
17	- Relational Model (RM) - Hierarchical (for XML) - Network Model (for ODBMS)					D 5.4 (RM)	
18	Relational Theory	Level 3					
19	- Relational Algebra	Level 2		3.2.2.6		D 5.4	
20	*Normalization*	Level 4				D 5.4	
21	Object-oriented Model	Level 5					
22	ODMG Standard	Level 5					
23	SQL Basics	Level 5		3.2.2.7			
24	QBE	Level 5					
25	Security	Level 3					
26	Transaction Processing	Level 4					
27	*Transaction Principle*	Level 4				D 5.6	

Figure 1. Mapping of DBTech Database Taxonomy to other CS curricula (partial view) [19]

of compromised transaction due to technological tricks like pooled connections or disconnected components.

To ensure problem solving competences beyond technical issues students have to develop their ability to work in teams, manage tasks, organise releases and orchestrate different versions. All these knowledge can be learned from real world projects.

E. *E-Learning Model*

We believe it is best to decide from the learning content, which learning concept will be best suited for a specific content. The e-learning model we present integrates different learning concepts (see Issing [6]):

- Learning as behavioral modification for *practical skills* and verification of the theory
- Learning as active information processing using assimilation and accommodation processes to build a mental model of the *theory*
- Learning as construction of knowledge used for problem based learning as in *project work*

All these concepts are used in an integrative way in order to get the most effective results in terms of applicable knowledge and profound cognition that enable abstraction and problem solving to a large extent. The design of the e-learning model (see Figure 2) starts with structuring the learning area guided by a taxonomy. The area is sliced with a minimum of dependencies and each chunk of learning content is represented in a theory unit, with examples and demonstrations of the theory. Hands-on experiments help to verify the theory. The global optimization task is to put together all aspects in balance with the target learning group.

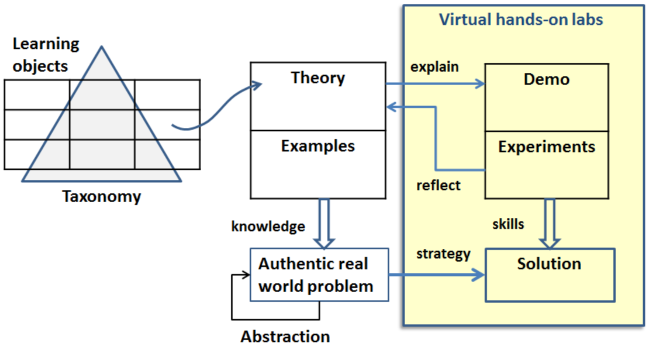


Figure 2. E-Learning Model Overview

Examples and demonstrations explain the theory, making it easier to understand. Hands-on experiments motivate and stimulate students to reflect the theory. Examples provide the students with analogous situations that could be applied and abstracted in the project work. The interrelation of all these elements provided in a virtual lab environment with the theory units and the examples are as shown in Figure 2.

The concrete real world problem forces the students to abstract from examples and construct a model of the problem world in order to find a solution.

IV. Technical Framework and Infrastructure

The framework of technologies provides a central, web based repository for teaching material, lab environments, multimedia, communication and collaboration tools.

A. E-Learning Portal

We provide all e-learning material through a portal (see http://dbtech.uom.gr and [22]) using Moodle as software platform. It contains all theory units, mostly as reading material, video lectures, tests, assessments and experimental lab environments that will be described in the following subsection. Local versions, like translations or modifications that fit the curriculum constraints are hosted and maintained at the project partners sites (https://relax.reutlingen-university.de for Reutlingen University, or https://elearn.haaga-helia.fi/moodle/login/index.php for Haaga-Helia University for Applied Sciences).

B. Virtual Laboratory Infrastructure

The lab environments are available either through technologies like desktop virtualization or virtual machines running computer software images. The latter is used when the image only uses free software. In this case, there is no need to control the number of downloads or to provide licences. After downloading the image it can run off-line. Free player for the image are available, e.g., VirtualBox.

For commercial software products where licenses are needed, the use of a desktop virtualization is more appropriate as it let easily control the number of remote application accesses. Citrix XenDesktop or VMware View are examples that provide a Virtual Desktop Infrastructure (VDI) for different operating systems.

VDI provides remote access to a pool of virtual machines through a connection broker. If the license policy is for a number of concurrent users it is no problem to limit the concurrent users with this software. Access control may be enforced by LDAP or Active Directory. The virtual machines are automatically managed in terms of multiple and customized instances of computer systems, applications, and for every users. Independent virtual machines may be assigned to avoid any resource access conflicts. Access to different operating systems is possible and the assignment to a client's PC may be persistent or transient.

As infrastructure for accessing the virtual machines from a client machine a local or public area network is needed. Client computers only need a web browser with ActiveX or Java Applet technology support. Such a support is given by the most common web browsers.

DBTech EXT uses a VDI operated by the University of Málaga. The number of concurrently active virtual machines depend on the resources (processor cores, memory, and disc space) provided. In the case of DBTech EXT labs Málaga uses two VMware servers with two quad-core processors and 32 Gigabytes of RAM each [23]. This infrastructure has enough power to run 96 concurrent virtual systems, each with 512 Megabytes of memory. The VDI architecture is presented in Figure 3 showing the VMware architecture consisting of a virtual center and two Hypervisor ESX servers that provide for multiple operating systems running on a single server. The broker is responsible for dispatching the connection requests from clients and to control the access with the help of an authentication service.

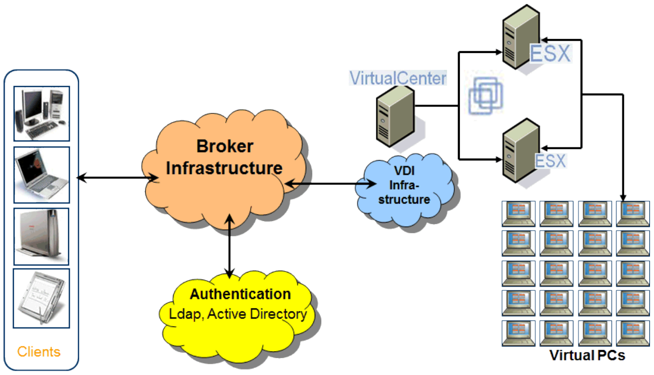


Figure 3. Virtual Desktop Infrastructure for virtual labs [23]

V. Experiences

The experiences mainly stem from two EU funded projects that were carried out during the years 2002-2005 and 2009-2010 (see http://www.dbtechnet.org). During the first project phase we identified important knowledge areas of *database systems* and syllabi of courses. The syllabi were later extended to a taxonomy and integrated within a unified learning concept.

A couple of example e-learning modules have been developed as testing material and these courses were used as teaching material for virtual workshops conducted during the second project. For the virtual workshop the e-learning platform was enhanced by communication and collaboration tools like Skype, discussion boards and upload areas for deliverables. Teaching material was structured and furnished with exercises and assignments for the students. The exercises used the previously described virtual infrastructure to guaranty a predefined and fully functional environment. Assessment of the student was done by on-line tests preferably in form of multiple choice questions.

The methodologies used to evaluate and assess our concept included informal and formal (survey conducted via the e-learning platform) feedback and self evaluation, discussions with students, and the results of written examinations (open and multiple choice questions).

The answers to the multiple choice questions were collected and assessed with the help of the e-learning system. However, the type of questions allowed only to test the analytical skills and not the construction of knowledge or innovative solutions. As consequence the project work was only assessed manually at the partner's institutions.

In Reutlingen study projects of real world problems are incorporated into the curriculum since more than 10 years. Over many generations of students the feedback was uniformly positive. Students appraise the real life character of the projects. In about one third of the projects, the problem was posed by a company that also collaborated with the students team. From the didactics point of view the motivation was kept high if the company or the university committed itself to use the project results. In most cases this was a software to be developed by the students.

Problem based learning confirmed the proposed high motivation if in addition the knowledge background of the project team was sufficient to master the problem. It was not necessary that each of the participants was an expert programmer or had managing competence. It was sufficient to have at least one with the necessary capability. In most cases this stimulated the team and resulted in an intensive team internal learning process. The supervising professor has the responsibility to make sure that the students with less knowledge will not become frustrated. The intervention could be additional training for the "weaker" students or to assign a different role to the "dominant" student. In individual situations we have been successful if the more knowledgeable student acts as a trainer for a while.

Comparing student teams that work physically together outperform teams that only work together virtually. In feedback discussions the students state a lower motivation and commitment to the project team if they worked remote without meeting each other. Asking for reasons the students named the missing personal contact and commitment. In contrast the teams that met regularly developed a culture of responsibility that supported motivation and contributed to the project success.

VI. Conclusion

The outstanding lessons learned of this long term e-learning experience can be summarized in three statements:

1) A key success factor is the adequate slicing of the knowledge domain. Only if this requirement is granted, the necessary small chunks of information are identified and can be prepared according to our e-learning model. If the chunks are not small and sufficiently independent it is hard to provide e-learning modules that can be worked through without the constant help of the teacher.
2) E-Learning is not superior to face-to-face teaching. It is more difficult to motivate the students. The preparation of study material is much more elaborate than for traditional teaching.
3) E-Learning scales better only for knowledge and comprehension level (Bloom's taxonomy) and partially for the application level. For higher level (deeper understanding) as synthesis, evaluation and analysis a stronger communication seems to be necessary for this cognitive levels.
   We found no way to automate the assessment of creative and constructive results like the assessment of a project or a software. This is a challenge for future work.

References

[1] Aristoteles, Eugen Rolfes, and Günther Bien, *Aristoteles: Philosophische Schriften 3: Nikomachische Ethik*, Translated by Eugen Rolfes, revised by Günther Bien, Meiner (1995)

[2] Johann Amos Comenius, *Didactica Magna*, 1657, In: Klaus Schaller (ed.), Ausgewählte Werke. vol. 1, Hildesheim: Olms, 1973, p. 82

[3] Gordon M. Burghardt, "The evolutionary origins of play revisited: Lessons from turtles". In M. Beckoff and J. A. Byers, (eds.), *Animal Play: Evolutionary, Comparative, and Ecological Perspectives*, Cambridge, UK: Cambridge Univ. Press, 1998

[4] Doris Bergen (ed.), *Play as a Medium for Learning and Development*, Portsmouth, NH: Heinemann, 1987

[5] Andreas Flitner, *Spielen - Lernen : Praxis und Deutung des Kinderspiels*, 3^{rd} ed., 1998

[6] Ludwig Issing, "Psychologische Grundlagen des Online-Lernens", in: Paul Klimsa and Ludwig Issing (eds.), *Online-Lernen - Handbuch für Wissenschaft und Praxis*, 2^{nd} ed., Oldenbourg V., München, 2011

[7] André Kukla, *Social Constructivism and the Philosophy of Science*, London: Routledge, 2000, ISBN 0415234190

[8] Benjamin Bloom, Max Engelhart, Walker Hill, and David Krathwohl, *Taxonomy of educational objectives: the classification of educational goals*; Handbook I: Cognitive Domain, New York, Longmans, Green, 1956.

[9] Hermann Astleitner, Iwan Pasuchin, and Christian Wiesner, "Multimedia und Motivation - Modelle der Motivationspsychologie als Grundlage für die didaktische Mediengestaltung", Medienpädagogik, Zeitschrift für Theorie und Praxis der Medienbildung, 22.3.2006, [Online] http://www.medienpaed.com/2006/astleitner0603.pdf, last access: 16.03.2012

[10] Hermann Astleitner and Christian Wiesner, "An Integrated Model of Multimedia Learning and Motivation". Journal of Educational Multimedia and Hypermedia, 2004, 13(1), 3-21. ISSN 1055-8896

[11] Manfred Prenzel, "Sechs Möglichkeiten, Lernende zu demotivieren", in: H. Gruber and A. Renkel (eds.), *Wege zum Können, Determinanten des Kompetenzerwerbs*, Bern, Huber, pp. 32-44, 1997

[12] Manuela Paechter, Karin Schweizer, and Bernd Weidenmann, "Lernen in virtuellen Seminaren: Neuigkeitsbonus oder Adaption an ungewohnte Lernbedingungen", in: F. Scheuermann (ed.), *Campus 2000 - Lernen in neuen Organisationsformen*, Münster, Waxmann, pp. 279-288, 2000

[13] Seymour Papert, *Constructionism: A New Opportunity for Elementary Science Education* , Massachusetts Institute of Technology, Media Laboratory, Epistemology and Learning Group, 1986,

[14] Ernst von Glasersfeld, ”Konstruktion der Wirklichkeit und der Begriff der Objektivität”; in: Heinz von Foerster: *Einführung in den Konstruktivismus*, publ. of Carl-Friedrich-von-Siemens-Stiftung, vol. 5; München: Piper, 1992; ISBN 3-492-11165-3

[15] Jochen Gerstenmaier and Heinz Mandl, ”Wissenserwerb unter konstruktivistischer Perspektive”, Zeitschrift für Pädagogik, 41. Jg, Nr. 6, pp. 867-888, 1995

[16] Thomas Connolly and Carolyn Begg, ”A Constructivist-Based Approach to Teaching Database Analysis and Design”, Journal of Information Systems Education (JISE), Vol. 17(1), pp. 43-53, 2006

[17] Christian Spannagel, *Benutzungsprozesse beim Lernen und Lehren mit Computern*, PhD-Thesis, Hildesheim, Franzbecker, 2007, ISBN-10: 3-88120-443-9

[18] Luis Fernández Sanz, ”Analysis of non technical skills for ICT profiles”, 5^{th} Iberian Conference on Information Systems and Technologies (CISTI), Santiago de Compostela, 2010

[19] Dimitris Dervos, Martti Laiho, and Fritz Laux, ”DBTech Pro Framework Reference Courses and Topics”, [Online] http://dbtech.uom.gr/mod/resource/view.php?id=592 (Excel-Format), last access: 16.03.2012

[20] Abraham Silberschatz, Henry Korth, and S. Sudarshan, *Database System Concepts*, New York, McGraw-Hill, 2011

[21] Gerhard Weikum and Gottfried Vossen, *Transactional Information Systems*, Morgan Kaufmann Publishers, 2002

[22] Georgios Evangelidis, Evangelos Pitsiougas, Dimitris Dervos, Martti Laiho, Fritz Laux, and José F. Aldana-Montes, ”DBTechNet portal: A Gateway to Education and Training for the European Database Technology Professional”, in Proceedings of the eRA-4 International Conference on the Contribution of I.T. to Science, Technology, Society and Education, 24-26 September 2009, Spetses, Greece.

[23] Ismael Navas Delgado, Maria del Mar Roldán García, José J. Farfán-Leiva, Martti Laiho, Fritz Laux, Dimitris A. Dervos, and José F. Aldana Montes, ”Innovative Unity in the Diversity of Information Management Skills Teaching and Training across Europe”, in IADAT Journal of Advanced Technology on Education, Vol. 3, No. 3, July 2009, pp. 407-445, ISSN: 1698-1073

Dynamic Composition of Curriculum for Computer Science Courses

Zona Kostić
zkostic@singidunum.ac.rs

Dragan Marković
dsamarkovic@singidunum.ac.rs

Aleksandar Jevremović
ajevremovic@singidunum.ac.rs

Ranko Popović
rpopovic@singidunum.ac.rs

Irina Branović
ibranovic@singidunum.ac.rs

Faculty of Informatics and Computing,
Singidunum University
Danijelova 32, Belgrade, Serbia

Abstract— **In this paper, a novel platform for curriculum development and design used for creating a Virtual University System is described. The platform has been developed based upon experience of using Web-based Computer Science virtual environment. The system consists of Interface, Automatic assessment and Tutoring modules, and is used for dynamic creation of new courses, syllabi, and curricula. The Web-based technology we applied enables the use of virtual environment for computer science courses in distance learning.**

Keywords - Dynamic curriculum; Virtual learning environments; Web-based education

I. Introduction

This paper describes a novel approach to curriculum development and design in creating Virtual University System. Most existing educational environments use Virtual Reality (VR) techniques, which concern the creation and use of the Virtual Learning Environment (VLE) [1]. Use of the VLE at universities requires collaboration and interaction between the onsite and distant students and instructors.

Distance learning is an alternative and a supplement to traditional classroom instruction. The interactive nature of virtual classroom addresses the main challenges found in distance education, namely student involvement and participation. Virtual classrooms often rely on software simulators of pedagogical computer systems. A number of software tools targeting teaching and learning in introductory courses in computer science have been proposed and developed [2]. Also, dynamic curriculum development has been an active area of research [3] [4].

Our research goal has been the development of Virtual University System based on the experience of applying online teaching and learning for the last decade. Our virtual learning environment is designed to address the limitations of existing distance education systems. The result of our work is practically implemented system for distance learning and dynamic creation of curricula, whose novelty with respect to all current educational systems is threefold:

- Enables integration of virtual and real classrooms;
- Applies a specific pedagogical approach based on students' feedback;
- Allows for dynamic creation of new syllabi and curricula.

A curriculum defines learning content of a course or program of study in terms of knowledge and skills, i.e., specifies main teaching, learning, and assessment methods. Curriculum also indicates learning resources required to support the effective delivery of the course. A syllabus describes the content of a specific program of study and can be thought of as a part of curriculum [3].

The elements of a curriculum are: learning outcomes, content, teaching and learning methods, assessment, and virtual learning resources. There are a number of papers dedicated to curriculum development and design, [5][6][7][8]. None of them offers specific pedagogical approach based on student profiles and educational materials retrieved by using specific module, taking into account previous teaching experience. The curricula are developed and modified dynamically, through monitoring and evaluating at the end of each semester, when instructors compare them with corresponding curricula of the most renowned universities. Starting point for development is an IEEE curriculum standard in which the existing programs are dynamically modified or simple changes made to individual lessons [9]. The instructor chooses course materials based on his preferences and student feedback.

Creating new curricula is modular, aiming to provide instructors with some of the background theory related to curriculum design and course development, while integrating best-practice approaches and recent trends in computer science education. Curriculum development is an iterative process in three phases: evaluation, content modification, adding new methods and corrections.

Our Virtual University System applies the Problem Based Learning (PBL) methodology by implementing it as project-based learning [10]. This methodology engages students to integrate theory and practice, and see the big picture, rather than only pieces of the problem.

The analysis of system use, together with the observed fact that students tend to spend much time in virtual reality, justify the choices taken in development of a virtual system and orientation towards integrating virtual environment with reality.

The rest of the work is organized as follows: Section 2 contains background and motivation, concerns with technologies, environments, platforms, and related work; Section 3 describes phases and modules in dynamic creation of curriculum; Section 4 gives a case study example of

Computer Graphics (CG); Section 5 describes the results of the application and evaluation; and finally, Section 6 concludes and presents future work.

II. BACKGROUND AND MOTIVATION

There are two essential parameters relevant for developing virtual learning environments: the first are technologies used for implementation, and the second are previously implemented virtual environments upon which new ones can be based. This section gives a short overview of 3D technologies and environments related to education.

A. 3D Technologies

There are many technologies which enable the presentation of 3D data on the Internet; an excellent survey is given in [11]. The most commonly used are X3D (eXtensible 3D) and WebGL (Web-based Graphics Library), both designed for the creation of interactive Web-based and broadcast-based 3D content, and suitable to integrate with multimedia. WebGL works without installing additional software, but only within a compatible Web browser. Regardless of the fact that X3D works at much lower level and needs installation of an appropriate plug-in, it works within any Web browser, and as scene-graph system and with XML encoding, it is much better choice for beginning students.

B. Environments

Learning Management Systems (LMS) dominate in e-learning; the most prominent examples are Blackboard [12], Moodle [13], ATutor [14], and dotLRN [15]. These are integrated systems which support a wide area of distance learning activities. Platforms are often used in education and they are commonly divided into commercial and non-commercial products. For example, Sloodle [16] integrates Moodle, the open source tool for learning with Second Life, the most used commercial platform. Both commercial and non-commercial platforms lack important functionalities, such as cooperativity, real-life experience, and desktop sharing, while offering only average graphics quality. Most Virtual Universities are based on commercial platforms.

C. Related Work

There are many systems conceptually similar to ours, but none of them offers full range of functionality. In particular, these systems are not extendible. For example, the virtual classroom with smart tutor described in [17] is a good solution, but only for single courses. There are also solutions using X3D for creating virtual learning classrooms and labs. EVE [18] is the closest solution to ours, especially when it comes to CG course, but without real-time streaming. Paper [19] describes an excellent course for engineering students, but without collaboration, text or video chat. The solution [20] combines many features of X3D and ActiveX in creating virtual lab, but lacks the possibility of dynamic curriculum creation. Finally, [21] offers an excellent interface with much functionality, but lacks groupwork support. With respect to all described educational systems, our approach is superior because of the integration of virtual and real classrooms, application of specific pedagogical approach based on students' feedback, and support for dynamic creation of new syllabi and curricula.

III. DYNAMIC CREATION OF CURRICULUM

Our Virtual University System consists of three software modules (Figure 1):

- Interface module
- Content retrieval module (MSearch)
- Assessment module (MTutor).

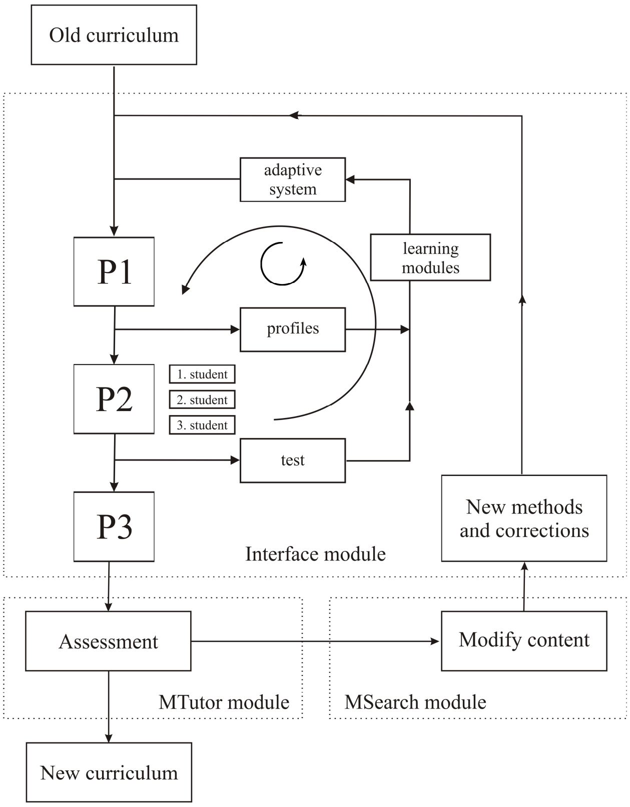


Figure 1. Three components of dynamic curriculum creation cycle.

Our platform allows for creating of new syllabi and curricula based on the previous ones. This process is dynamic; it iteratively uses student profiles and educational materials retrieved by using specific module.

We will begin by describing separate functionalities of three different modules, and proceed with an overview of dynamic curriculum creation process.

A. Interface module

This module defines the interaction between participants in the process, and serves as the starting point for assessment. Interface module is based on the following components:

- Component which defines student’s profile by using learning and adaptive modules;
- Component which defines the connection between a student and other group members working on the same project, teaching assistant and instructor;
- Component which defines the connection between instructors and teaching assistants.

Interface is not an independent component, but instead incorporates feedback from other modules. Specifically, the system is able to dynamically change the course contents based on student profile; instructor tracks the changes and selectively incorporates them into the syllabus based on his own judgment.

B. Modeling of virtual environment

Since most of computer science courses are taught in computer labs, the necessary component of our virtual environment is a virtual model of a real laboratory. A virtual lab is a component of interface module.

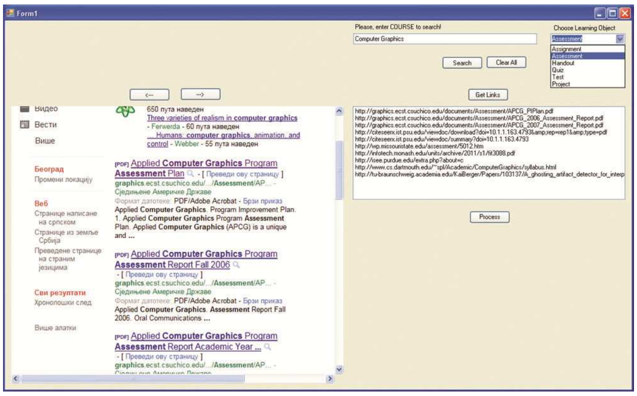
The basic components of any laboratory are virtual client computers. The physical interface takes the form of a classroom equipped with thin clients. By virtualizing client computers, important benefits such as flexibility and availability are achieved. Virtual machines are much cheaper and easier to install/clone than physical ones. Furthermore, they consume less electrical power and space. Also, because different labs are executed on core servers at different times, virtualization allows hardware reuse by switching between different virtual machines. The default protocol for accessing client virtual machines is VNC, making them platform independent. Software configuration of each virtual machine in a lab is different, customized for course requirements.

Other components of the lab include electronic educational materials (documents, simulators, and evaluation systems) and real-time streams. The laboratory core is built using cluster servers which provide execution of virtual machines, configured for educational processes.

C. Content module

MSearch retrieves educational material (in the form of pdf, ppt, and html files) from different universities’ Web sites based on user query for a text phrase and a query for an image name. MSearch allows personalizing the learning process, as well as reusing previous research efforts, results, and experience. At the end of each semester, teaching and testing materials are compared with other universities. This approach enables instructors to assess the difficulty level of their lessons and practical work with respect to the same courses at other universities, and to include other instructors’ experience in new syllabi and curricula.

MSearch searches Google Web and Google Images, with a query filtering set for academic domains. Dodget Get Links displays top 10 links for a text query and top 10 links for an image query (Figure 2). The top three ranked links (or a link of the user's choice) are passed to the crawler. Afterwards, complete indexing, retrieving images, as well as doc, rtf, and pdf documents from selected Web pages is performed. Indexing of documents is multimodal, i.e., both text and image metadata are indexed.



Please, enter COURSE to search!

Computer Graphics

Choose Learning Object

Assessment

Assignment
Assessment
Handout
Quiz
Test
Project

650 пута наведен
Three varieties of realism in computer graphics - Ferwerda - 60 пута наведен
... Humans: computer graphics, animation, and control - Webber - 55 пута наведен

[PDF] Applied Computer Graphics Program Assessment Plan - [Преведи ову страницу]
graphics.ecst.csuchico.edu/.../Assessment/AP... - Сједињене Америчке Државе
Формат датотеке: PDF/Adobe Acrobat - Брзи приказ
Applied Computer Graphics. Program Improvement Plan. 1. Applied Computer Graphics Program Assessment Plan. Applied Computer Graphics (APCG) is a unique and ...

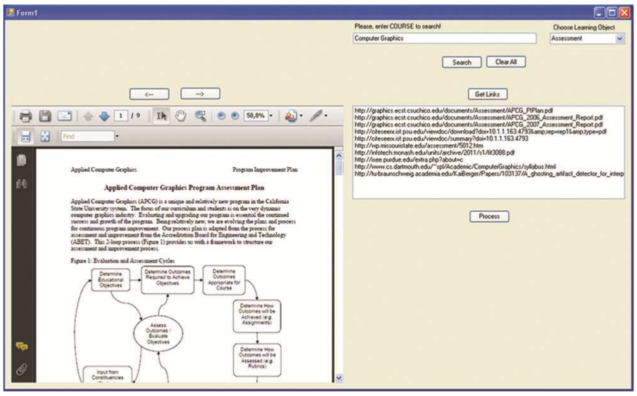
[PDF] Applied Computer Graphics Program Assessment Report Fall 2006 - [Преведи ову страницу]
graphics.ecst.csuchico.edu/.../Assessment/AP... - Сједињене Америчке Државе
Формат датотеке: PDF/Adobe Acrobat - Брзи приказ
Applied Computer Graphics. Assessment Report Fall 2006. Oral Communications ...

[PDF] Applied Computer Graphics Program Assessment Report Academic Year ... - [Преведи ову страницу]
graphics.ecst.csuchico.edu/.../Assessment/AP... -

```
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_PIPlan.pdf
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_2006_Assessment_Report.pdf
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_2007_Assessment_Report.pdf
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.4793&amp;rep=rep1&amp;type=pdf
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.4793
http://wp.missouristate.edu/assessment/5012.htm
http://infotech.monash.edu/units/archive/2011/s1/fit3088.pdf
http://isee.purdue.edu/extra.php?about=c
http://www.cs.dartmouth.edu/~spl/Academic/ComputerGraphics/syllabus.html
http://tu-braunschweig.academia.edu/KaiBerger/Papers/103137/A_ghosting_artifact_detector_for_interp
```

Figure 2. Text query and choosing learning object.

Based on the results that MSearch presents, the instructor is able to decide which link to use to compare retrieved exam questions with his own (Figure 3). The instructor can also define new questions or decide to import the questions retrieved from the selected link into the assessment module. This approach enables instructor to define the difficulty level of questions, and by taking into account students’ feedback he can balance students’ load accordingly. At the same time, the instructor incorporates changes of the curricula at universities that he considers relevant. For example, Figures 2 and 3 illustrate the use of materials from Berkeley, which is among top computer science universities.


Form1
Search
Clear All
<--
-->
Get Links
Process
1 / 9
58,8%
Find
Determine Educational Objectives
Determine Outcomes Required to Achieve Objectives
Determine Outcomes Appropriate for Course
Determine How Outcomes will be Achieved (e.g. Assignments)
Determine How Outcomes will be Assessed (e.g. Rubrics)
Assess Outcomes / Evaluate Objectives
Input from Constituencies

Please, enter COURSE to search!

Computer Graphics

Choose Learning Object

Assessment

Applied Computer Graphics Program Improvement Plan

Applied Computer Graphics Program Assessment Plan

Applied Computer Graphics (APCG) is a unique and relatively new program in the California State University system. The focus of our curriculum and students is on the very dynamic computer graphics industry. Evaluating and upgrading our program is essential the continued success and growth of the program. Being relatively new, we are evolving the plans and process for continuous program improvement. Our process plan is adapted from the process for assessment and improvement from the Accreditation Board for Engineering and Technology (ABET). This 2-loop process (Figure 1) provides us with a framework to structure our assessment and improvement process.

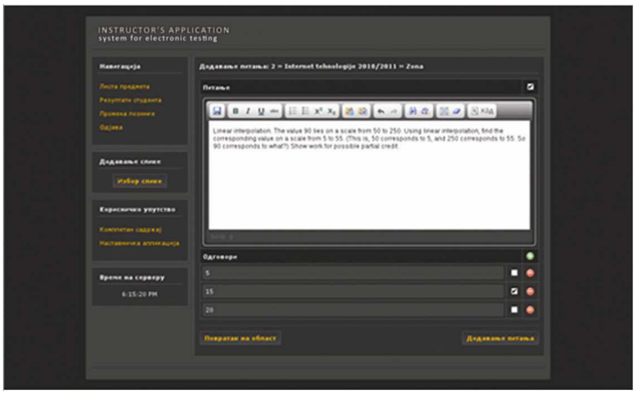
Figure 1: Evaluation and Assessment Cycles

```
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_PIPlan.pdf
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_2006_Assessment_Report.pdf
http://graphics.ecst.csuchico.edu/documents/Assessment/APCG_2007_Assessment_Report.pdf
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.4793&amp;rep=rep1&amp;type=pdf
http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.163.4793
http://wp.missouristate.edu/assessment/5012.htm
http://infotech.monash.edu/units/archive/2011/s1/fit3088.pdf
http://isee.purdue.edu/extra.php?about=c
http://www.cs.dartmouth.edu/~spl/Academic/ComputerGraphics/syllabus.html
http://tu-braunschweig.academia.edu/KaiBerger/Papers/103137/A_ghosting_artifact_detector_for_interp
```

Figure 3. Text query results.

D. Assessment module

MTutor, is subsequently used for computer testing and defining appropriate learning steps by applying student-centered rules during learning and testing. MTutor system allows using textual and multimedia querying with true/false and multiple choice questions. Statistical processing of students' results is enabled on class, single student, question and answer base levels. Besides grading students, these results are important for other purposes, such as recognizing high-quality questions (that are correlated with overall course goals) and eliminating low-quality questions. In this way, course questions pool is iteratively improved after every examination, and as a result overall validity and reliability of grading method used is improved. Another purpose is comparing student groups and class results by using integrated T-test component. This feature enables instructors to measure the impact of curriculum changes that were applied. Finally, statistical processing of students' results enables instructors to quantitatively measure performance of a single student during the course.



Додавање питања: 2 = Internet tehnologije 2010/2011 = Zona

Питање

Linear interpolation. The value 90 lies on a scale from 50 to 250. Using linear interpolation, find the corresponding value on a scale from 5 to 55. (This is, 50 corresponds to 5, and 250 corresponds to 55. So 90 corresponds to what?) Show work for possible partial credit.

Одговори

5 □
15 ☑
20 □

Figure 4. An example question in MTutor module.

The process starts and finishes with searching for various test types and comparing them with the existing ones using the automatic search engine and finishes by gathering the selected questions into the system for the final exam (MTutor module).

IV. PROCESS OF CREATING DYNAMIC CURRICULUM

Heterogeneous groups of students are formed based on students' learning styles. In general, adaptive systems for individual learning include modeled entities on which the decisions on adaptation are based (i.e., user preferences) described in [22].

Dynamic curriculum is developed in the following three phases (Figure 1):

- Phase 1: Pretesting, using the adaptive system and self-assessment to obtain student profiles and change the learning modules. After creating learning modules tailored to a specific student, heterogeneous groups of students are formed. A project is assigned to each group.
- Phase 2: The three subprojects are merged into the "big picture", followed by interactive assessment (supervised by an instructor) and exams, a final step in which everyone participates. The instructor makes final suggestions based on students' opinions about the work of others. The results of the second phase are documentation of the final project, with an option to use different types of assessments, and adaptive testing.
- Phase 3: Objective assessment, using students' educational materials, is done as a final exam, under the responsibility of a supervisor. Student evaluation is based on this assessment. Upon conclusion of the exam, the supervisor corrects submitted materials and stores them. The supervisor and/or groups of students, based on the previous experience, proceed in making changes to learning modules of the old syllabus and iteratively change it.

Described phases are performed by using resources of a virtual lab, while the instructor is allowed to control and eventually alter their execution.

Course design and curriculum development tools are saved to be reused for subsequent iterations of the same process. The process of creating a dynamic syllabus is repeated for many courses and is eventually used for constructing a new curriculum. We use a pedagogical approach based on student-virtual system-instructor feedback. Each module contains an introduction to the topic along with explanations of concepts, illustrations, and, where appropriate, interactive components to visualize algorithms related to that topic.

Example: COMPUTER GRAPHICS LAB

We will explain the previously described ideas on an example in which creating a computer graphics lab is assigned as a group project to students.

IEEE curriculum [7] defines modeling, visualization, and virtual reality as principal interrelated fields of interest for a computer graphics course. These three main components were used for defining a computer graphics group project. The project is implemented through three subprojects assigned to each group; the subprojects are:

- Modeling 3D virtual laboratory environment (adding objects such as tables, PCs, routers etc.)
- Enabling communication by using predefined chat and real-time streaming components, and also connecting components created in the first subproject
- Integrating the two previously created components into a Web-based application.

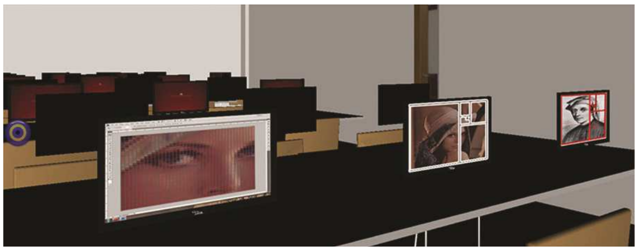


Figure 5. Implementation of the 3D lab for computer graphics course.

The first subproject consists of creating a 3D lab model in X3D with objects for computer graphics. Many tools for

3D modeling and animation enable exporting in X3D format; students used 3ds Max software for creating a final model of the lab (Figure 5).

The second subproject involves integrating virtual client computers into the graphics laboratory. The physical interface consists of thin clients. Web interface to client virtual machines is implemented as an X3D lab, with integrated Virtual Network Connection (VNC) client software. The idea is that 3D environment should replicate real laboratory.

Further development of the virtual lab is done in X3D-Edit by adding functionalities, such as interaction, animation, and navigation. Animation involves adding timers and interpolators to drive continuous events, interaction concerns mouse-based picking and dragging, while navigation defines user movements, collision, and visibility detection. It is also necessary to enable collaboration (synchronous communication) within X3D lab. A lab must contain a link to dynamic libraries, described in an XML file. A student must use Flash to connect to Virtual Machines.

V. THE RESULTS OF THE APPLICATION AND EVALUATION

In this section, we present a general evaluation of our virtual educational model and its implementation during two semesters. Qualitative and quantitative evaluation of the proposed educational approach was conducted. The qualitative evaluation included a number of student surveys and discussions with instructors. Surveys attempted to determine what students perceived as good educational tools and how they assessed the overall effectiveness of this approach. Students mainly complained about the steep learning curve of the system, and their suggestions have been implemented in subsequent iterations of our system. Few lessons at the beginning of each course were dedicated to training students on how to use the system. Also, manuals and tutorials explaining the toughest steps of using VUS were developed and integrated into the system. Manuals also describe how to use the existing tools in all different phases of learning, self testing, group work, and exam taking.

Students also complained that during preparation for self testing cannot decide which tools to use. This was also perceived as a difficulty in learning process; as a remedy, the option to choose the right tool based on experience of previous users (i.e., statistical data on most used tools and materials) was added. The statistical data are regularly updated upon completion of the course and used in a new syllabus.

Yet another students' complaint was the low speed of access to the environment. It had been noticed that in days before the exam the number of VUS site visitors had rose up to 80% of students. The problem of the access speed to the environment was solved by using the appropriate software for load balancing.

We also measured system efficiency by comparing final results of students who used the laboratories through the virtual and physical interface.

The exam was conducted in a controlled environment (Computer Graphics exam, Figure 5.) with 40 students, of which 20 used the physical environment, and the other 20 used the virtual environment. The testing process was supervised by two instructors and two teaching assistants.

The results of the statistical analysis (T-test) of final exam test scores are presented in Table 1.

TABLE I. RESULTS OF THE STATISTICAL ANALYSIS OF FINAL EXAM TEST SCORES ON A SAMPLE OF 40 STUDENTS DIVIDED INTO TWO EQUALLYSIZED GROUPS

	Control group (physical environment)	**Experimental group (virtual environment)**
No of students	20	20
Mean result (points)	71.55	79.35
Standard deviation	17.73	15.20
Variance	314.35	232.04

Comparing results with the corresponding values in the T-table (for a statistically acceptable p - value of 0.05, calculated on the basis of degrees of freedom for both groups) showed no statistically significant differences in results achieved by control and experimental groups.

The results of this statistical analysis, combined with the observed fact that students tend to spend much more time engaged in virtual reality than in using multimedia instructional materials, further encourage the orientation towards a completely virtual environment in the future.

The results of our analyses prove the usability of our approach and justify the attempts to dynamically adapt new curricula by comparing the results achieved by different generations of students.

VI. CONCLUSION AND FUTURE WORK

Our Virtual University System is developed as a distance education system to enable fully integrated real and virtual labs. The labs posses unique characteristics that address the limitations of existing distance education systems. New courses and their syllabi are defined using three software modules. The new Web-based curricula are created iteratively, taking into account previous teaching experience. The results of our experiments are promising enough to encourage further integration of Virtual University System with sensors, semantic 3D model retrieval, and 3D searching. In the next development phase, a significant research effort will be put into automating all phases in creating a Virtual University System by introducing artificial intelligence, through curriculum sequencing as a way of helping the student to find the best route through the educational material. The plan for future research is to integrate students' profiles as input parameters for curriculum sequencing to provide personalized learning paths through the course content (texts, exercises, examples, and questions).

ACKNOWLEDGMENTS

The work presented here has been supported by the Serbian Ministry of Education and Science (project grant no. III44006)

REFERENCES

[1] C. G. Burdea and P. Coiffet, “Virtual reality technology”, 2nd Edition, John Wiley & Sons, pp. 464, 2003.

[2] K. E. Sanders and R. McCartney, “Program assessment tools in computer science: a report from the trenches”, Proceedings of the 34th SIGCSE technical symposium on Computer science education (SIGCSE '03). ACM, pp. 31-35, 2003.

[3] J. McKimm, “Curriculum design and development”, School of Medicine, Imperial College Centre for Educational Development, pp. 32, 2007.

[4] J. Brewer, A. Harriger, and J. Mendonca, "Beyond the Model: Building an Effective and Dynamic IT Curriculum", Journal of Information Technology Education, pp. 441-458, 2006.

[5] T. Groover and J. Kabara, “The design and implementation of a pre-college computer science curriculum for underrepresented high school students”, Frontiers In Education Conference: Knowledge Without Borders, Opportunities Without Passports, pp. T3A-22 - T3A-23, 2007.

[6] J. Chookittikul and W. Chookittikul, "Six sigma quality improvement methods for creating and revising computer science degree programs and curricula," 38th Annual Frontiers in Education Conference, pp. F2E-15-F2E-20, 2008.

[7] The Development of the IEEE/ACM Software Engineering Curricula

http://www.enel.ucalgary.ca/People/yingxu/Publications/Papers/IEEE%20CR50_Wang.pdf (accessed April 6, 2012).

[8] K. Georgouli, "Virtual Learning Environments-An Overview," 15th Panhellenic Conference on Informatics, pp. 63 – 67, 2011.

[9] Computer Science Curriculum 2008

http://www.acm.org/education/curricula/ComputerScience2008.pdf (accessed April 6, 2012).

[10] G. Simic and A. Jevremovic, “Problem-based learning in formal and informal learning environments”, Interactive Learning Environments,

DOI: 10.1080/10494820.2010.486685, 2010.

[11] B. Turonova, „3D Web Technologies and Their Usability for The Project 3D Mobile Internet“, Technical Report, Research and Development Center for Mobile Applications, Faculty of Electrical Engineering, Czech Technical University in Prague, pp. 18, 2009.

[12] L. Ling and H. Lie, "Construction of Web-delivery Elaborate Courses Based on Blackboard: As an Example to the Course for Basic Circuit Analysis," International Forum on Information Technology and Applications, pp.287-289, 2010.

[13] M. Amelung, K. Krieger, and D. Rösner, "E-Assessment as a Service," IEEE Transactions on Learning Technologies, pp. 162-174, 2011.

[14] V. Gonzalez-Barbone, M. Llamas-Nistal, "eAssessment: Trends in content reuse and standardization," 37th Annual Frontiers In Education Conference - Global Engineering: Knowledge Without Borders, Opportunities Without Passports, pp.T1G-11-T1G-16, 2007.

[15] D. Huerva, J. Velez, and S. Baldiris, “Adaption of Courses and Learning Environment to the User Context in dotLRN”, Proceedings of the 2008 International Conference on Computational Intelligence for Modelling Control & Automation (CIMCA '08), pp. 1264-1267, 2008.

[16] Z. Guomin and Z. Jianxin, "An Educational Value Analysis of SLOODLE-Based Distributed Virtual Learning System," Second International Workshop on Education Technology and Computer Science, pp. 402-405, 2010.

[17] Y. Hu and G. Zhao, “Virtual Classroom with Intelligent Virtual Tutor”, Proceedings of the 2010 International Conference on e-Education, e-Business, e-Management and e-Learning, pp.34-38, 2010.

[18] C. Bouras, C. Tegos, V. Triglianos, and T. Tsiatsos, “X3D Multi-user Virtual Environment Platform for Collaborative Spatial Design”, Proceedings of the 27th International Conference on Distributed Computing Systems Workshops, pp. 40 – 40, 2007.

[19] P. Goeser, W. Johnson, F. Hamza-Lup, and D. Schaefer, “VIEW: A Virtual Interactive Web-based Learning Environment for Engineering,” IEEE Advances in Engineering Education Journal, Special Issue on Research on e-Learning in Engineering Education, pp. 24, 2011.

[20] S. Liang and P. Rong-jiang, “A 3D virtual experiment framework based on X3D and ActiveXL”, 2010 International Conference on Audio Language and Image Processing, pp. 1035–1039, 2010.

[21] Haiqing, Y. Guofu, and F. Jie., “Research on the Collaborative Virtual Products Development Based on Web and X3D”, Proceedings of the 16th International Conference on Artificial Reality and Telexistence, pp.141-144, 2006.

[22] S. Marković, N. Jovanović, R. Popović, and A. Jevremović, "Adaptive Distance Learning and Testing System", Computer Applications in Engineering Education, DOI: 10.1002/cae. 20510, 2010.

A system to help students analyze errors in their programs by supporting programming comprehension in assembly programming exercises

Yuichiro Tateiwa
Graduate School of Engineering
Nagoya Institute of Technology
Nagoya, Japan
tateiwa@nitech.ac.jp

Daisuke Yamamoto
daisuke@nitech.ac.jp

Naohisa Takahashi
naohisa@nitech.ac.jp

***Abstract*—There are several students who give up exercises because they cannot specify their program errors to fix. We considered the reason were the following. One is that the students do not have enough comprehension of their programs – questions ask their understanding of control structure, computer resource control, and behavior. Another is that procedures to specify program errors are complex because an assembly program has a lot of instructions. Furthermore, oversight, which is caused by misunderstanding questions and checking a lot of items, is also the causes. The purpose of this study is to develop a system which generates expression for specifying program errors by helping students understand their program comprehension. The features on realizing the system are making use of chunks, dynamic backward slices, and correct answer samples. We conclude that the expression is helpful to specify program errors according to an evaluation experience.**

Keywords-programming; assembly language; program slice; chunk;

I. Introduction

The “Systems Programming” course offered by the Department of Computer Science, Nagoya Institute of Technology, aims to help students understand hardware activities that occur in response to application software requests for computer resource control (e.g., controlling registers and main memory) and control structures (sequencing, selection, iteration, and function/procedure). Therefore, the class includes an assembly programming exercise in which students translate high-level-language (e.g., C) programs, whose activities are regarded as application software requests, into low-level-language programs (e.g., assembly language CASLL-II [1], whose activities are regarded as hardware activities.

In class, students solve exercises on structured programming using the above-mentioned control structure. Questions in the exercise include the requirements of program behavior, computer resource control, and control structure in the form of text and C programs. Students’ answers (hereafter, “answer program”) are considered correct if they do not contain all of the following error types.

・Behavior error: answer program does not behave according to the requirements of questions.

・Computer resource control error: computer resources are not used according to the requirements of questions.

・Control structure error: control structure is not designed according to the requirements of questions.

We developed a programming exercise system that can automatically detect these error types [2]. This enables students to immediately confirm whether their answers are correct. First, the system obtains two detailed “trace data” (i.e., a sequence of pairs of instructions that are executed stepwise and the computer resources that are updated by the execution) by the stepwise execution of the answer program and the correct answer program (hereafter, simply “correct program”) with test cases. Each pair of trace data for a given step in the sequence is called “step data”. Next, the system extracts characteristic points (e.g., order of label appearance, variable values, relation between locations of instructions) from both trace data. If differences are detected between corresponding characteristic points, the system judges that the answer includes errors. After the evaluation is completed, the results are displayed to the student; if errors are present, the test case is also displayed. A student whose answers are incorrect is expected to try to specify the causes of errors by using her/his program, question, and test case. S/he should analyze control structures, trace the dependency of computer resource control, simulate the test case behavior, and so on. S/he should consider whether these satisfy the requirements of the question. If not, s/he should specify error instructions in her/his program.

However, several students are unable to complete these tasks. We believe that some students lack an understanding of the control structure, computer resource control, and program behavior (we collectively term this as “program comprehension”). Others find the procedures for specifying causes of errors too complex because an assembly program contains many instructions. Finally, some problems are due to oversight stemming from, say, misunderstanding of program requirements and the need to check many items.

In this study, we develop a function that generates expressions that help students specify causes for their errors by supporting their program comprehension. We call these

ISBN: 978-1-61208-200-4

expressions “assistance expressions.” The function detects errors in answer programs and classifies them as 1) control structure errors or 2) behavior errors and computer resource control errors, and it generates assistance expressions for each error.

In order to implement this function, we use “chunks,” “dynamic backward slices,” and correct answers. A “chunk” is a meaningful block (sequential elements). It is easier to understand and memorize programs when they are expressed as a sequence of chunks. A “dynamic backward slice,” a type of program slice, is a sequence of instructions that influences variable v, which is defined at time r when an instruction is executed, when executing a program with input arguments x. Hereafter, we call time r the “execution point” and the triple of x, r, and v, the “slicing criterion (x, r, v).” The dynamic backward slice at an execution point where variables differ between the answer and the correct program includes the causes of errors (inclusion of error instructions and lack of necessary instructions). Therefore, its use can help students specify the causes of errors.

The function generates the following three types of assistance expressions.

(1) Chunk expression of programs: It is important to read and understand programs in order to specify the causes of errors. However, as mentioned above, assembly programs are difficult to read and understand. We solve this problem by explicitly expressing control structures in a program and meaningful instruction sequences in control structures by using chunks. A chunk containing instructions is called a “static chunk.” Expressing programs using static chunks helps in understanding control constructions and specifying causes of errors, in addition to reading programs. For example, although instructions of “readout arguments” are connoted by the control structure “sequence,” it is possible to show a composition of the control structure “sequence” to students by defining a static chunk “readout arguments.” In addition, it is possible to show the control structures that contain errors to students by defining a static chunk of error implementations. This function generates an expression that is a static chunk sequence of both an answer and a correct program, and the expression includes instructions that constitute the chunks of the answer program. The aim of the expression is to help students notice the differences in control structures and specify the causative instructions.

(2) Chunk expression of trace data: Trace data is useful for specifying the causes of behavior error and computer resource control errors. However, it is difficult to read and understand because it is large in amount, and it is troublesome to match to a program because of the use of different expressions. We solve this problem by expressing trace data using chunks that are related to static chunks. Hereafter, a chunk containing trace data is called a “dynamic chunk.” This function generates an expression that is a dynamic chunk sequence of both an answer and a correct program to help students notice the difference between the two.

(3) Projection expression of error steps: Some instructions, called as an “error instruction sequence,” contain important clues for specifying the causes of program behavior error and computer resource control errors, and their execution results influence the difference between the trace data of an answer and a correct program. A student specifies the causes of errors in the answer program by confirming the relationships between the error instruction sequence and the remaining instructions and by comparing the execution results of the answer and the correct program. However, (1) and (2) are not suitable for such procedures. The former does not show error instruction sequences and their execution results to a student, and therefore, it is difficult to compare an execution result between the answer and the correct program. The latter does not show all instructions of the answer program, and therefore, it is difficult to confirm the relationships between instructions in the answer program. Accordingly, we try to solve the problems by computing an error instruction sequence and expressing it and its execution result for the answer program. The function generates an expression that is based on (1) with an error instruction sequence of an answer program, the execution results of it and a correct program.

II. RELATED WORKS

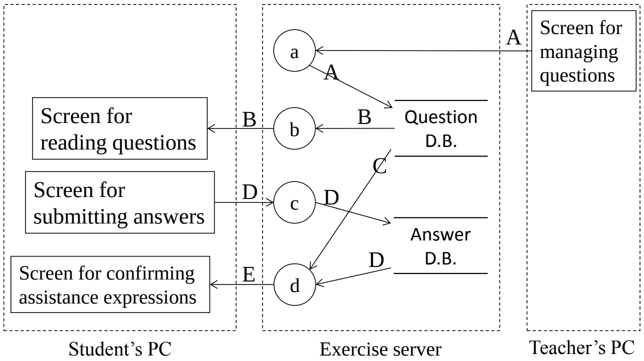
A “program slice” is a set of instructions that influences a certain instruction in a program [3][4][5]. Program slices are used for program debugging and program comprehension. A dynamic backward slice in this study is characterized by its slicing criterion which is a point that causes difference of program behavior and computer resource control between an answer and a correct program. Namely, its feature is to use not only an answer program but also a correct program. Using this criterion, we can compute a slice that includes the causes of errors.

Static chunks are calculated by pattern matching between a program and a pattern that defines a rule of a static chunk. As a related study that uses pattern matching in assembly programs, W.Kozaczynski et al. proposed the replacement of frequently used instruction sequences with simple expressions and comments for easy readability and understandability [6]. We, however, propose a method to generate information that simplifies the reading and understanding of programs and trace data and the correspondence between programs and trace data using static chunks.

III. FUNCTION FOR GENERATING ASSISTANCE EXPRESSIONS

A. Placement in our programming exercise system

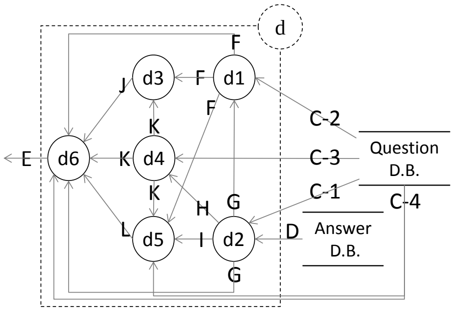
Fig. 1 shows the structure of our system. The exercise system consists of an exercise server on a machine, and web browsers for each student and teacher on their PCs. And the machine and the PCs are connected to the Internet. A student receives a question from the exercise server (“b”), composes an answer to the question, and submits it to the server (“c”). The server detects errors in the answer, and generates assistance expressions that depends on the error type by using the answer program, a correct program, test cases, “static chunk conditions,” and “evaluation item sets” (“d”). A “static chunk condition” is a condition for extracting


Screen for managing questions
A
a
A
Screen for reading questions
B
b
B
Question D.B.
C
Screen for submitting answers
D
c
D
Answer D.B.
D
Screen for confirming assistance expressions
E
d
Student's PC
Exercise server
Teacher's PC

a: Register questions
b: Get questions
c: Submit answers
d: Generate assistance expressions

A: questions, correct programs, test cases, static chunk conditions, evaluation item sets
B: questions
C: 1) correct programs, 2) test cases, 3) static chunk conditions, 4) evaluation item sets
D: answers programs
E: assistance expressions

Figure 1. System structure



C-1, C-2, C-3, C-4, D, E: refer to Fig. 1
F: step data of answer programs and correct programs
G: converted answer programs and converted correct programs
H: converted answer programs and correct programs, descriptors for routine bodies of answer programs and correct programs
I: converted answer program
J: descriptors for dynamic chunks of answer programs and correct programs
K: types and descriptors of static chunks in answer programs, types and descriptors of static chunks in correct programs
L: descriptors for error instruction sequences
d1: generate step data
d2: convert programs and extract routine bodies
d3: extract dynamic chunks
d4: extract static chunks
d5: extract error instruction sequences
d6: generate assistance expressions

Figure 2. Structure of function for generating assistance expressions

instructions from a program. An "evaluation item set" is a set of input/output variables in static chunks that are compared between the answer and the correct program. The student confirms the true/false judgment of her/his answer; if the answer is false, s/he corrects it on a Screen using assistance expressions. A teacher registers questions, correct programs, test cases, static chunk conditions, and evaluation item sets with the exercise server before the student starts the exercises ("a").

B. Function structure

Fig. 2 shows the structure of the function for generating assistance expressions in "d" of Fig. 1. "d2" extracts instructions of answer and correct programs and converts their formats to that shown in Fig. 3. The function then extracts bodies of "routines" from both converted programs (Section III.C.1). Henceforth, a "routine" is a main routine or a subroutine that is a sequence of instructions from "start" to "end." In addition, the body of a routine (called a routine body) is a sequence that consists of machine instructions and macro instructions. "d1" generates each step data by stepwise execution of an answer and a correct program with test cases (Section III.C.2). "d4" extracts static chunks of an answer and a correct program by comparing instructions of a routine body with static chunk conditions (Section III.C.3). "d3" extracts dynamic chunks by comparing stepwise executed instructions with instructions that consist of static chunks (Section III.C.4). "d5" compares input/output variables between an answer and a correct program in the order of executed instructions. In the comparison, the variables are selected in accordance with an evaluation item set, and their values are computed using step data. When "d5" detects a difference in the comparison, it computes a dynamic backward slice of the variable that causes the difference, and it regards the slice as an error instruction sequence (Section III.C.5). "d6" classifies an answer program as being correct or as containing a control structure error or behavior/computer resource control error, and it generates assistance expressions (Section III.C.6).

Hereafter, this paper describes a new data type using a structure in the C language. However, "struct" is omitted in the member and variable declarations. For example, a data structure X with int type members a and b is described as "struct X {int a; int b;}." A variable z of data type X is described as "X z;." A member a of z can be referred to by "z.a."

C. Implementation methods

1) Convert programs and extract routine bodies

Answer and correct programs conform to the CASL-II grammar. We have extended this grammar by adding operation codes SSP and LSP, which save and load a stack pointer, respectively. These are necessary for the class to implement a general procedure of a function call, which is implemented by stack frame operations in most assembly languages such as GNU assembly language.

The data structure of a program in our algorithms is a character array. Instructions in a program are converted from their original formats into the one shown in Fig. 3, and then stored in the character array (e.g., Fig. 4). Henceforth, "<>" indicates that the elements therein can be omitted, and "{}" indicates that elements therein are necessary. We call

```
<label>{space}{operation code}<{space}{operand1}<,operand2><,operand3>>▯ n
```

Figure 3. Format of an instruction in our algorithms

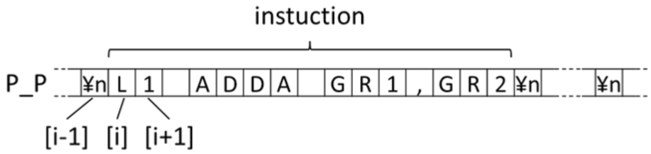


Figure 4. An example of a variable P_P that contains a program

character strings that are located at the label, operation code, operand1, operand2, and operand3 identifiers. Variables P_P and A_P store the answer program and correct program, respectively.

A routine is a main routine or a subroutine that is a sequence of instructions from start to end. In addition, the body of a routine (called a routine body) is a sequence that consists of machine instructions and macro instructions. Routine bodies are extracted from converted answer and correct programs. The data structure of a routine body in our algorithms is of the Block type. Block type is designed for pointing to a sub array "sub_array" in an array "array," and it is defined as "struct Block{int s; int e;}." Members s and e are the indexes of elements in "array" that are respectively the first and last elements of "sub_array." P_R and A_R, which are Block-type array variables, respectively store the routine bodies of the answer program and correct program in the order found in the programs. For example, a character string that is from P_P[P_R[0].s] to P_P[P_R[0].e] is the instruction sequence of the first routine body in P_P (Fig. 5).

2) Generate step data

A step datum is a result that is generated by stepwise executing a program with a test case as input. It consists of an executed instruction, names of computer resources that are updated by executing the instruction, and their values. The data structure of a computer resource in our algorithms is defined as "struct CmpRes{char[] n;char[] v;}"; members n and v are the first addresses of character arrays that stores a computer resource name and computer resource value, respectively . The data structure of a step datum is defined as "struct Step{int i;CmpRes[] cr;}"; member cr is the first address of a list that stores computer resources that are updated by stepwise execution on an instruction whose first character is the i+1st character in a program. Step type arrays P_S and A_S respectively store step data of the answer and the correct program in order of stepwise execution. For example, P_S[k] is a step datum of the k+1th stepwise execution in answer program P_P. P_P[P_S[k].i] is the first character that is an executed instruction; the character string whose first address is pointed to by P_S[k].cr[0].n is a computer resource name that is the first updated by the executed instruction; and P_S[k].cr[0].v points to the character string of its value (Fig. 6).

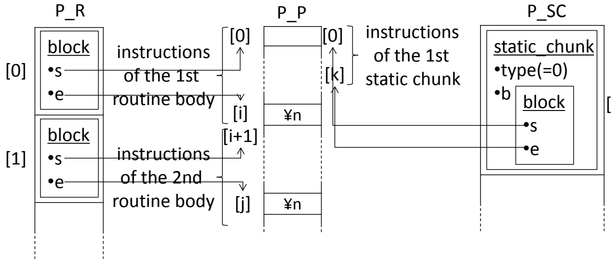

P_R
block
•s
•e
[0]
instructions of the 1st routine body
block
•s
•e
[1]
instructions of the 2nd routine body
[0]
[i]
[i+1]
[j]
P_P
¥n
¥n
[0]
[k]
instructions of the 1st static chunk
P_SC
static_chunk
•type(=0)
•b
block
•s
•e
[0]

Figure 5. An example of the arrays P_P, P_R, and P_SC

3) Extract static chunks

As mentioned in Section I, static chunks help in reading/understanding programs, understanding control structures, and specifying causes of control structure errors. The control structure in the class includes the sequence, selection, repetition, and function/procedure. To extract other static chunks, it is necessary to define new static chunk conditions and register them in the system.

A static chunk is an instruction sequence that has a meaning in toto. The data structure of a static chunk in our algorithms is defined as "struct SChunk{int t; struct Block b;}"; member t is a type (Tabel 1), and member b is a character string that is from the b.s+1th character to the b.e+1th character, and this character string is an instruction sequence that constitutes this static chunk. SChunk type arrays P_SC and A_SC respectively store static chunks of an answer and a correct program in the order of extraction. For example, a character string that is from P_P[P_SC[i].b.s] to P_P[P_SC[i].b.e] is an instruction sequence that consists of the i+1th static chunk that is extracted from an answer program. P_SC[0] in Fig. 5 is extracted first; its type is a sequence according to P_SC[0].type=0, and it consists of a character string that is from P_P[P_SC[0].b.s] to P_P[P_SC[0].b.e].

The data structure of a static chunk condition in our algorithms is a pair of a static chunk type and a condition for extracting instruction sequences (called the "instruction sequence condition"). Because the identifier and number of instructions depend on the questions, it is difficult to define instruction sequence conditions using only instructions of CASL-II; it is necessary to define many conditions. Regular expressions and pattern matching are effective ways to solve this problem. However, it is necessary to consider the following points.

Requirement 1: It is necessary to extract instruction sequences that include arbitrary and same identifiers at multiple points of the sequence. For example, in an

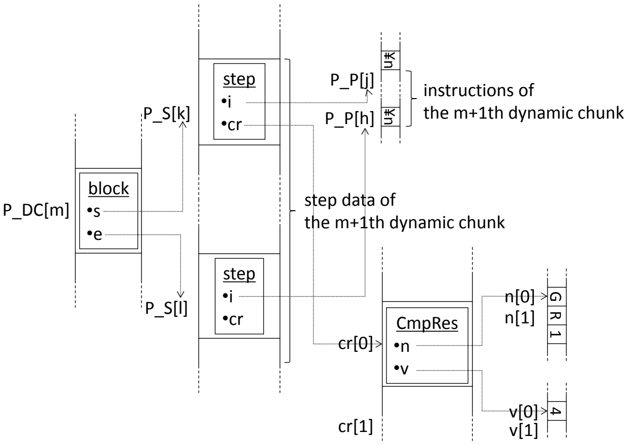


Figure 6. An example of the arrays P_S and P_DC

instruction sequence selection, the operand of the operation code branch and label of the branch destination are an arbitrary and same identifier.

Requirement 2: It is necessary to extract a character string that is matched to a particular part in a regular expression. This is used for specifying a character string, extraction target, by the character before and after it. For example, Fig. 7-a) shows that the instruction sequence from line 1–4 is the chunk selection. The tail of the chunk is characterized by an instruction immediately in front of the branch destination (line 5). In such cases, the first 5 lines are extracted, and then, line 5 is removed.

We adopted Perl, whose regular expressions enable a character string extracted by a regular expression to be referred to from the back of the expression itself (for requirement 1). It is possible to refer to a character string that is matched to a group by group numbers after completing the matching (for requirement 2). Therefore, an instruction sequence condition consists of a regular expression and a group number.

Fig. 7-b) shows an example of an instruction sequence condition for the chunk selection. It expresses 1 line character string by dividing it into multiple lines owing to space limitations. The first group is a regular expression from the left parenthesis in line 1 to the right parenthesis in line 4, and it is designed to extract an instruction sequence of a chunk selection. The second group is “(\w+)” in line 2, and it is referred to by "\2" in lines 3 and 5. When it applies this regular expression to the instruction sequence in Fig. 7-a), because line 2 in Fig. 7-a) matches the line 2 in Fig. 7-b), a character string that matches the second group is considered as “L1,” and “\2” in lines 3 and 5 is specified as “L1.” In addition, the character string from lines 1–4 in Fig. 7-a) matches the first group, and it is extracted as the instruction sequence of a chunk selection.

The data structure of a static chunk condition in our algorithms is defined as “struct ChunkCond{int t;char[] ptn; int g;}”; member t is a type of a static chunk (Table 1), member p is the first address of a character string of an instruction sequence, and member g is a group number for designating a character string in the extraction. A ChunkCond type array CC stores static chunk conditions.

Fig. 8 shows an algorithm for extracting static chunks. The program in the class conforms to the rule of structured programming, and it is not allowed to jump between routines by operation code jump. Therefore, we developed a simple algorithm that extracts static chunks from every routine because there is no chunk through multiple routines. A function “int N (T[] array1)” returns the number of elements in arbitrary type T array array1. A function “void merge(T[] array1, T[] array2)” merges the elements of arbitrary type T arrays array1 and array2 so that the elements of array2 are appended to the end of array1. A function “void add(T[] array1, T elem1)” appends arbitrary type T elem1 to the end of arbitrary type T array array1. When P_P and P_R, or A_P and A_R, are respectively stored in P and R and static chunk conditions are stored in CC, following which static_chunk(P, R, CC) is executed, an SChunk type variable that contains static chunks is obtained.

TABLE I. STATIC CHUNKS

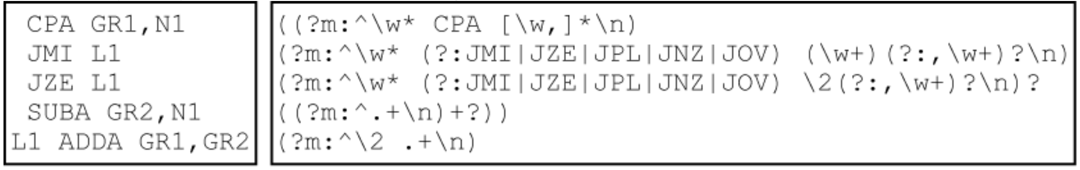
type	chunk name
0	sequence processing
1	start processing of function
2	start processing of function (with errors)
3	end processing of function
4	end processing of function (with errors)
5	readout processing of arguments
6	readout processing of arguments (with errors)
7	repetition processing
8	selection processing

4) Extract dynamic chunks

A dynamic chunk is a sequence of step data that is generated by single stepwise execution of instructions from the start to the end of a static chunk. The data structure of a dynamic chunk in our algorithms is Block type; members s and e in P_S and A_S indicate that step data from P_S[s] to P_S[e] or A_S[s] to A_S[e] is included in the extracted dynamic chunk. Dynamic chunks of an answer and a correct program are stored in Block-type arrays P_DC and A_DC, respectively, in order of extraction. For example, a Step-type array from P_S[P_DC[m].s] to P_S[P_DC[m].e] is a step data instruction of the m+1th dynamic chunks that are extracted. Fig. 9 shows an algorithm for extracting dynamic chunks. When P_SC and P_S, or A_SC and A_S, are respectively stored in SC and S, following which dynamic_chunk(SC, S) is executed, a Block-type array that contains dynamic chunks is obtained.

5) Extract error instruction sequences

An “error instruction sequence” is a sequence of instructions that affects the difference between the trace data of an answer and a correct program. We call such instructions “candidates for error instructions.” Error instruction sequences and their execution results are important clues for specifying the causes of program behavior errors and computer resource control errors. Trace data of an answer and a correct program are compared based on evaluation item sets. An evaluation item set is a designation of computer resources that are compared



```
 CPA GR1,N1
 JMI L1
 JZE L1
 SUBA GR2,N1
L1 ADDA GR1,GR2
```

```
((?m:^\w* CPA [\w,]*\n)
(?m:^\w* (?:JMI|JZE|JPL|JNZ|JOV) (\w+)(?:,\w+)?\n)
(?m:^\w* (?:JMI|JZE|JPL|JNZ|JOV) \2(?:,\w+)?\n)?
((?m:^.+\n)+?))
(?m:^\2 .+\n)
```

a) Instructions

b) Instruction sequence condition

Figure 7. An example of instructions and an instruction sequence condition for selection processing

ISBN: 978-1-61208-200-4

```
SChunk[] static_chunk(char[] P,Block[] R, ChunkCond[] CC){
  SChunk SC[0];
  for(int i=0;i<N(R);i++)
    merge(SC, f(P,R[i],CC,0));
  return SC;}
SChunk[] f(char[] P,Block b,ChunkCond[] CC,int cc_i){
  SChunk SC[0];
  if(cc_i == N(CC)){
    SChunk sc = {0, b};
    add(SC, sc);
    return SC;
  }
  Block m = match(P,b,CC[cc_i]);
  if(m.s == -1 && m.e == -1)
    merge(SC,f(P,b,CC,cc_i+1));
  SChunk sc = {CC[cc_i].t, m};
  add(SC, sc);
    Block next = {m.e+1, b.e};
    Block prev = {b.s, m.s-1};
      if(b.s == m.s && m.e < b.e){
        merge(SC,f(P,next,CC,cc_i));
      }else if(s < m.s && m.e < e){
        merge(SC,f(P,prev,CC,cc_i+1));
        merge(SC,f(P,next,CC,cc_i));
      }else if(s < m.s && m.e == e){
        merge(SC,f(P,prev,CC,cc_i+1));
      }
      return SC;}
Block match(char[] P, Block b, ChunkCond cc){
      Block m = {-1,-1};
    Extract character string ss that matches the cc.g-th group in
cc.ptn from the character string that is from P[b.s] to P[b.e];
    if (ss exists)
    m.s=x and m.e=y on the condition that ss is the character
string that is from P[x] to P[y], and b.s<=x and y<=b.e
    return m;}
```

Figure 8. An algorithm for extracting static chunks

between an answer and a correct program in the input or output.

The data structure of an error instruction sequence in our algorithms is a Block-type array. An element e of the array indicates a candidate for error instruction, which is a character string from the e.s+1th to the e.e+1th character in a program. An error instruction sequence is stored in the order extracted from a dynamic backward slice. The data structure of an evaluation item set in our algorithms is a char-type four-dimensional array. CK is designed for holding evaluation item sets; CK[a][0][c] points to the first address of the c+1th variable name that is compared between an answer and a correct program in the input of the a+1th static chunk that is extracted, such as P_SC[a] and A_SC[a]. CK[a][1][c] points to the first address of the c+1th variable name that is compared between an answer and a correct program in the output of the a+1th static chunk that is extracted, such as P_SC[a] and A_SC[a].

```
Block[] dynamic_chunk(SChunk[] SC,Step[] S){
  Block DC[0];
  for(int j=0;j<N(SC);j++){
    for(int i=0;i<N(S);i++){
      if(S[i].i==SC[j].b.s){
        Block dc = {i, -1};
        while((i+1<N(S)) &&
             (S[i].i<S[i+1].i) &&
             (S[i+1].i<=boi(SC[j].b.e))){
          i++;
        }
        dc.e = i;
        add(DC, make_array(dc));
      }}}
  return DC;
}
```

Figure 9. An algorithm for extracting dynamic chunks

Fig. 10 shows an algorithm for extracting an error instruction sequence. A function “error_ins” first judges, in the order of executing instructions, whether an answer

```
Block[] error_ins(char[] P_P, char[] A_P, SChunk[]
P_SC,SChunk[] A_SC,Step[] P_S,Step[] A_S,char [][][][] CK){
for(int i=0,k=0;i<N(A_S)||k<N(P_S);){
  Using j and l, which are A_S[i].i == A_SC[j].b.s and P_S[k].i
  == P_SC[l].b.s, for(int m=0;m<N(CK[j][0][m];m++){
     if(value(P_S,k-1,CK[j][0][m]) differs from value(A_S,i-
     1,CK[j][0][m])){
        int s = def(P_S,k-1,CK[j][0][m]);
        if(s!=-1){
           b[w]={x,y} on the condition that the instruction that is
           from P_P[x] to P_P[y] is equal to the w+1th instruction
           in a dynamic backward slice on slicing criterion=(s,
           CK[j][0][m], the test case);
           return b;
        }else{
           b[w]={x,y} on the condition that the instruction that is
           from P_P[x] to P_P[y] is equal to the w+1th instruction
           from the first instruction in P_SC[l];
           return b;
  }}}
  Using j and l, which are A_S[i].i == boi(A_P,A_SC[j].b.e) and
  P_S[k].i == boi(P_P,P_SC[l].b.e), for(int m=0;
  m<N(CK[j][1][m]; m++){
     if(value(P_S,k-1,CK[j][1][m]) differs from value(A_S,i-
     1,CK[j][1][m])){
        int s = def(P_S,k-1,CK[j][1][m]);
        if(s!=-1){
           b[w]={x,y} on the condition that the instruction that is
           from P_P[x] to P_P[y] is equal to the w+1th instruction
           in a dynamic backward slice on slicing criterion=(s,
           CK[j][1][m], the test case);
           return b;
         }else{
           b[w]={x,y} on the condition that the instruction that is
           from P_P[x] to P_P[y] is equal to the w-th instruction
           from the last instruction in P_SC[l];
           return b;
  }}}
  if(i<N(P_DC))  i++;
  if(j<N(A_DC))  j++;
}}
int def(Step[] S,int i,char[] n){
  for(;0<=i;i--)
     for(int j=0;j<N(S[i].CR);j++)
        if(the character string that is pointed to by S[i].CR[j].n is
        equal to the character string that is pointed to by n)
        return i;
 return -1;}
 char[] value(Step[] S,int i,char[] n){
   int j=def(S,i,n);
   if(j!=-1)
     for(int k=0;k<N(S[j].CR);k++)
        if(the character string that is pointed to by S[j].CR[k].n is
        equal to the character string that is pointed to by n)
        return S[j].CR[k].v;
 return "";}
```

Figure 10. An algorithm for extracting an error instruction

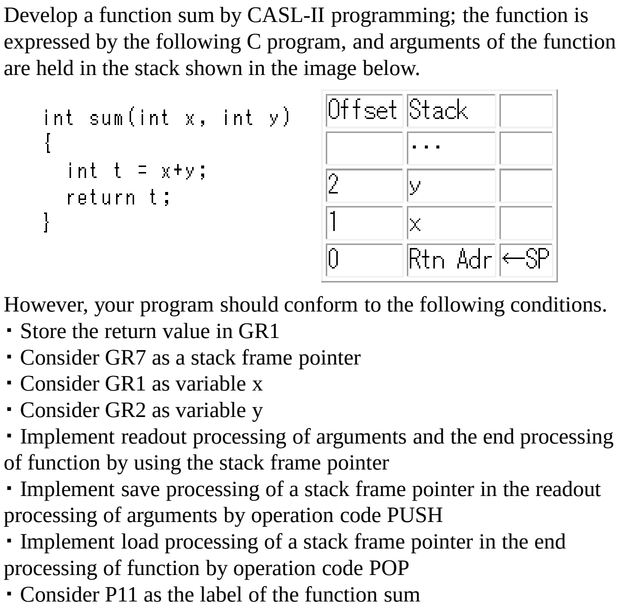
program is equal to a correct program in terms of the input and output, which are designated in CK. Lines 3–5 search the first step data of a static chunk in the order of executing instructions, and lines 6–7 compare variables, which are designated in an evaluation item set, between an answer and a correct program at the input point that is immediately before the stepwise execution of the first instruction of a static chunk. In addition, lines 3 and 21–23 search the end step of the static chunk in the order of executing instructions, and lines 24–25 compare variables, which are designated in an evaluation item set, between an answer and a correct program at the output point that is immediately after the stepwise execution of the end instruction of a static chunk. Lines 10–13 and 28–31 compute a dynamic backward slice when the compared computer resource at the input or output point differs between the answer and the correct program, and the compared resource is defined at the comparison time. Otherwise, when the compared resource is not defined at the comparison time, lines 16–18 and 34–36 consider instructions that are from the instruction at the comparison time to the first instruction of a program as an error instruction sequence. When instructions from P[x] to P[y] include P[e], a function "int boi(char[] P, int e)" returns x. When error_ins(P_P,P_SC,A_SC, P_S, A_S, CK) is executed, a Block-type array that contains an error instruction sequence is obtained.

6) Generate assistance expressions

Our system first tries to detect a control structure error. If such errors are not detected, next, it tries to detect a behavior error and a computer resource control error. A control structure error is detected when a static chunk sequence of an answer program differs from that of a correct program. If such an error is detected, the system generates a chunk expression of the program by using static chunks of the answer and the correct program, and the answer program. A chunk expression of the program places static chunk sequences of the answer and the correct program side-by-side with the instructions of the answer program. A behavior error and a computer resource control error are detected when an error instruction sequence contains instructions. When such an error is detected, the system generates a chunk expression of trace data by using the answer program and dynamic chunks of the answer and the correct program. A chunk expression of trace data places dynamic chunk sequences of the answer and the correct program side-by-side. In addition, the system generates a projection expression of error steps by using the answer program, evaluation item sets, static chunks of the answer and the correct program, and step data of the answer and the correct program. A projection expression of error steps adds a chunk expression of the program to an error instruction sequence of the answer program and its execution results, and the execution result of a correct program.

IV. Prototype System

This system holds the following static chunk conditions; sequence processing, start processing of function, start processing of function (with errors), end processing of function, end processing of function (with errors), readout processing of arguments, readout processing of arguments (with errors), repetition processing, selection processing. The question shown in Fig. 11 requires the implementation of a function sum that adds two arguments and stores the result in GR1. The behavior of the function in assembly is specified by a question sentence and a C program. For computer resource control, the use of all registers in the assembly program is specified by the corresponding C program. For example, the use of GR1 is specified by "assign GR1 to variable x" in the question sentence and variable x in the C program, which is assigned to the first argument.



Develop a function sum by CASL-II programming; the function is expressed by the following C program, and arguments of the function are held in the stack shown in the image below.

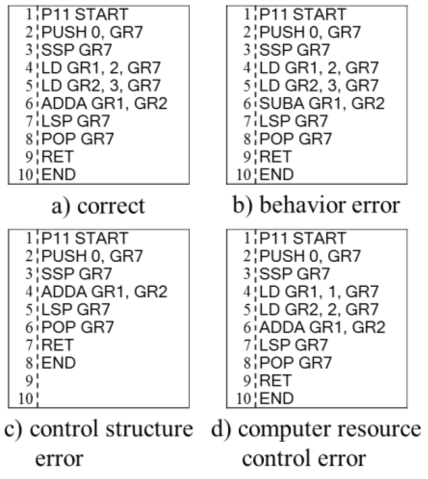
```
int sum(int x, int y)
{
  int t = x+y;
  return t;
}
```

Offset	Stack	
	···	
2	y	
1	x	
0	Rtn Adr	←SP

However, your program should conform to the following conditions.
・Store the return value in GR1
・Consider GR7 as a stack frame pointer
・Consider GR1 as variable x
・Consider GR2 as variable y
・Implement readout processing of arguments and the end processing of function by using the stack frame pointer
・Implement save processing of a stack frame pointer in the readout processing of arguments by operation code PUSH
・Implement load processing of a stack frame pointer in the end processing of function by operation code POP
・Consider P11 as the label of the function sum

Figure 11. A question on function implementation

Fig. 12 shows a correct answer (a) for the question shown in Fig. 11 and three incorrect answers (b, c, and d). In terms of behavior error, addition is correct instead of subtraction at line 6. In terms of control structure error, the instructions



```
 1 P11 START
 2 PUSH 0, GR7
 3 SSP GR7
 4 LD GR1, 2, GR7
 5 LD GR2, 3, GR7
 6 ADDA GR1, GR2
 7 LSP GR7
 8 POP GR7
 9 RET
10 END
```

a) correct

```
 1 P11 START
 2 PUSH 0, GR7
 3 SSP GR7
 4 LD GR1, 2, GR7
 5 LD GR2, 3, GR7
 6 SUBA GR1, GR2
 7 LSP GR7
 8 POP GR7
 9 RET
10 END
```

b) behavior error

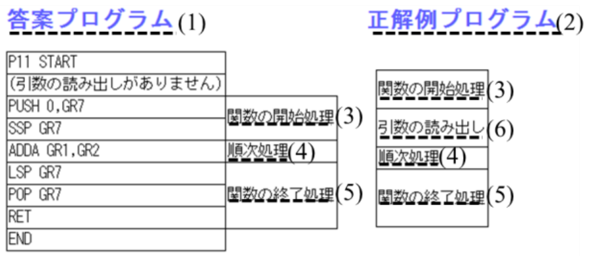
```
 1 P11 START
 2 PUSH 0, GR7
 3 SSP GR7
 4 ADDA GR1, GR2
 5 LSP GR7
 6 POP GR7
 7 RET
 8 END
 9
10
```

c) control structure error

```
 1 P11 START
 2 PUSH 0, GR7
 3 SSP GR7
 4 LD GR1, 1, GR7
 5 LD GR2, 2, GR7
 6 ADDA GR1, GR2
 7 LSP GR7
 8 POP GR7
 9 RET
10 END
```

d) computer resource control error

Figure 12. A correct answer and incorrect answers of the question in Fig. 11



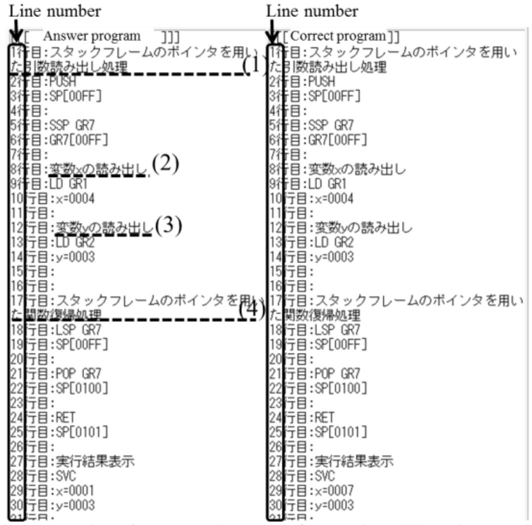
(1): answer program (2): correct program
(3): start processing of function
(4): sequence processing
(5): end processing of function
(6): readout processing of arguments

Figure 13. Chunk expression of program

described at lines 4 and 5 in the answer program are missing. The two instructions are used for readout of the arguments. In terms of computer resource error, the control structure is correct but the second operands of LD at lines 4 and 5, which are used for the readout of arguments, are incorrect.

Fig. 13 shows an assist expression for an answer program that contains a construction structure error (Fig. 12-c). A chunk "readout processing of arguments" is not shown because the answer program lacks instructions for the readout of arguments. A student notices this by comparing the chunk sequences of the answer and the correct paper.

Fig. 14 shows trace data of the correct and the answer program containing a behavior error (Fig. 12-b). A student specifies the causes of errors by confirming the difference between the trace data of his/her program and the correct program. Line 1 in Fig. 14 indicates the start of the readout of the argument. Lines 8 and 12 indicate the processing. Line 17 indicates the start of the end of the processing of the function. Furthermore, the student can notice the difference between the variables in his/her program and the correct program at line 29. The student can notice that lines 1–5 in his/her program in Fig. 12-b are correct because lines 1–14 in the chunk expression of the trace data of his/her program and the correct program are identical. Additionally, the student can notice that lines 7–9 in his/her program in Fig. 12-b are correct because lines 17–25 in the chunk expression of the trace data of his/her program in Fig. 14 and the correct program are identical. Therefore, the student can understand that the error in his/her program is caused at line 6 in Fig. 12-b.



```
[ Answer program ]]]
1行目:スタックフレームのポインタを用いた引数読み出し処理
2行目:PUSH
3行目:SP[00FF]
4行目:
5行目:SSP GR7
6行目:GR7[00FF]
7行目:
8行目:変数xの読み出し
9行目:LD GR1
10行目:x=0004
11行目:
12行目:変数yの読み出し
13行目:LD GR2
14行目:y=0003
15行目:
16行目:
17行目:スタックフレームのポインタを用いた関数復帰処理
18行目:LSP GR7
19行目:SP[00FF]
20行目:
21行目:POP GR7
22行目:SP[0100]
23行目:
24行目:RET
25行目:SP[0101]
26行目:
27行目:実行結果表示
28行目:SVC
29行目:x=0001
30行目:y=0003
```

```
[Correct program]]
1行目:スタックフレームのポインタを用いた引数読み出し処理
2行目:PUSH
3行目:SP[00FF]
4行目:
5行目:SSP GR7
6行目:GR7[00FF]
7行目:
8行目:変数xの読み出し
9行目:LD GR1
10行目:x=0004
11行目:
12行目:変数yの読み出し
13行目:LD GR2
14行目:y=0003
15行目:
16行目:
17行目:スタックフレームのポインタを用いた関数復帰処理
18行目:LSP GR7
19行目:SP[00FF]
20行目:
21行目:POP GR7
22行目:SP[0100]
23行目:
24行目:RET
25行目:SP[0101]
26行目:
27行目:実行結果表示
28行目:SVC
29行目:x=0007
30行目:y=0003
```

(1): readout processing of arguments by a stack frame pointer (2): readout a variable x
(3): readout a variable y (4): end processing of function by a stack frame pointer

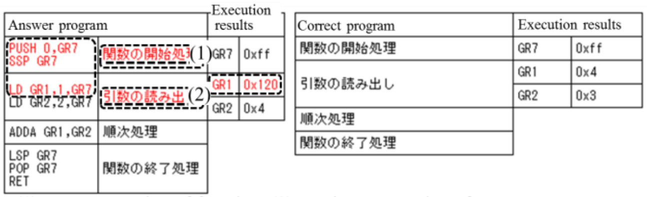
Figure 14. Chunk expression of trace data

Fig. 15 shows the projection expression of error steps for the computer resource control error in Fig. 12-d. The instructions in the dashed rectangle include causes of errors, and they are a dynamic backward slice that is computed because the value of GR1 of the answer program differs from that of the correct program at the output point of the second chunk. GR7 of the answer program is equal to that of the correct program at the output point of the first chunk, but GR1 differs at the output point of the second chunk. Therefore, the causes are narrowed to the following; in the second chunk, the instructions for GR7 are missing, and the instructions for GR1 are missing or incorrect. In a similar manner, narrowing down instructions that need to be reviewed using a dynamic backward slice and showing instructions and their execution results can help students to specify the causes of errors.

V. EVALUATION EXPERIMENT

An experiment was carried out to evaluate the effectiveness of assistance expressions. Chunk expressions of programs (expression 1), chunk expressions of trace data (expression 2), and projection expression of error steps (expression 3) help students specify the causes of their errors. We conducted a questionnaire survey for our system in the "Systems Programming" class, which is for second-year students in our university. The class includes four assembly programming exercises, and the students attempted 13 questions using our system. After finishing the fourth exercise, we distributed the questionnaires, which were aimed at determining how effective our system was in helping students specify the causes of their errors. The subjects answered the questions using the following five-point scale: 5 - very much, 4 - a lot, 3 - somewhat, 2 - not much, 1 - not at all. The questionnaires also contained space for comments.

24 valid responses were obtained. The average is rounded off to three decimal places, and the p-value is the value computed by a Wilcoxon test. The average and p-value of expression 1 are 3.88 and 0.0007, those of expression 2 are



Answer program		Execution results	
PUSH 0,GR7 SSP GR7	関数の開始処(1)	GR7	0xff
LD GR1,1,GR7 LD GR2,2,GR7	引数の読み出(2)	GR1	0x120
		GR2	0x4
ADDA GR1,GR2	順次処理		
LSP GR7 POP GR7 RET	関数の終了処理		

Correct program	Execution results	
関数の開始処理	GR7	0xff
引数の読み出し	GR1	0x4
	GR2	0x3
順次処理		
関数の終了処理		

(1): start processing of function (2): readout processing of arguments

Figure 15. Projection expression of error steps

4.25 and 0.0002, and those of expression 3 are 4.29 and below 0.0001. The averages and the p-values confirm that all suggested expressions helped the subjects specify the causes of their errors. The comment for expression 2 is “I could notice errors in my program behavior using the expression, but I spent a lot of time specifying the corresponding instructions in my program.” Because expression 3 is designed to resolve such problems, we will link expressions 2 and 3 using a hyperlink in future work. The comment for expression 3 is “It is helpful to narrow down instructions that are related to errors.” On the other hand, “This expression was not very helpful to specify operand order errors.” Such a solution is beyond the scope of our study at this time. We will develop functions to solve such problems in future work.

VI. CONCLUSION AND SUMMARY

In this study, we proposed a system that generates expressions for helping students specify causes of errors by helping them comprehend their program; students can use the developed functions to automatically judge whether their programs are correct. We developed this system because some students are unable to perform the above mentioned tasks in assembly programming exercises. In this system answers, correct answers, test cases, evaluation item set, and static chunk conditions are provided as inputs, and true-false judgments, chunk expressions of programs, chunk expressions of trace data, and projection expressions of error steps are provided as outputs. To generate such expressions, we suggested extraction methods for static chunks, dynamic chunks, and error instruction sequences. Through a questionnaire survey, we evaluated the effectiveness of the suggested expressions in helping students specify the causes of their errors. The results suggested that the expressions were quite helpful.

REFERENCES

[1] IPA, JITEC, “Information Technology Engineers Examination,” http://www.jitec.ipa.go.jp/1_13download/shiken_yougo_ver2_1.pdf (in Japanese), pp. 3-8, 2011.

[2] K. Miyati, N. Takahashi, “Implementation and Evaluation of a Computer-Aided Assembly Programming Exercise System with a Function of Structural Anomaly Detection,” IEICE TRANSACTIONS on Information and Systems (in Japanese), Vol.J91-D, No.02, pp. 280-292, 2008.

[3] Mark Weiser, “Programmers Use Slices When Debugging,” Communications of the ACM, Vol. 25, no. 7, pp. 446-452, 1982.

[4] H. Agrawal, J. Horgan, “Dynamic Program Slicing,” SIGPLAN Notices, Vol.25, No.6, pp. 246-256, 1990.

[5] Tankut Akgul, Vincent J. Mooney III, Santosh Pande, “A Fast Assembly Level Reverse Execution Method via Dynamic Slicing,” Proceedings of ICSE, pp. 522-531, 2004.

[6] W.Kozaczynski, E.S.Liongosari, J.Q.Ning, “BAL/SRW : Assembler re-engineering workbench. Information and Software Technology,” Vol.33, no.9, pp. 675-684, 1991.

Mashing up the Learning Environment

Evaluating a widget-based approach to Personal Learning Environments

Fredrik Paulsson
Interactive Media and Learning (TUV/IML)
Umeå University
Umeå, Sweden
fredrik.paulsson@edusci.umu.se

Abstract — **Different types of Virtual Learning Environments (VLE) have evolved and there is a steady ongoing progression of different concepts. During the last 10-15 years Learning Management Systems have dominated. Learning Management Systems are often presented as the solution for a range of educational needs. This paper presents a study of a mashup approach to the VLE using web widgets. A prototype was developed and discussed, covering technological aspects such as modularity, integration and adaptability as well as some pedagogical aspects, such as pedagogical flexibility and technological responsiveness. An alternative modular approach to the implementation of VLEs is suggested based on recent developments within web technology, stressing the use of standards and simplicity in order to address common problems of complexity and inflexibility resulting in poor conformance to pedagogical requirements.**

Keywords – LMS; MUPPLE; web widgets; mashup; VLE; PLE; e-learning.

I. Introduction

Different types of *Virtual Learning Environments (VLE)* have evolved over the years and there is a steady ongoing change and progression of different ideas and concepts for the VLE. During the last 10-15 years much has revolved around concepts Learning Platforms, such as *Learning Management Systems* (LMS). These systems are often presented as a common solution for a range of educational needs – much like a “Business System” for learning and education. However, the LMS have been criticized for being too inflexible and hard to adapt to different pedagogical contexts and needs (see, e.g., [1], [2] and [3]). The LMS are also criticized for having too much focus on the administrative aspects of learning with little support for pedagogical activities and pedagogical processes. Hence, having a strong focus on *Learning Management* rather than on actual learning and pedagogical activities per se - as the name actually suggests. From a system perspective LMS are commonly criticized for being designed and implemented in a silo-like fashion, contributing to lock-in effects of information and processes - very similar to the critique that is often heard about business systems in general. There is also a built-in conflict between the development and implementation of systems like LMS on the one hand and the development of social software and Web 2.0 on the other hand. While many LMS that are currently in use try to create a well-defined kind of “shielded community” for learning, web 2.0 is associated with open communities, global social interaction and open information services that can be used as building blocks for new services - such as for a *Personal Learning Environment* (PLE). However, observe that the notion of services for Web 2.0 refer to services that targets users and are not equivalent to services as in Service Oriented Architectures (SOA), which is to be regarded as a software design paradigm [4]. While the technology platform underlying Web 2.0 services may very well be a SOA platform, there is an unfortunate mix-up of those two rather different notions of services when discussing Web 2.0.

In order for services to be used as building blocks in such compositions (i.e., a mashup) the building blocks need to be well defined and with well-defined interfaces. Many web 2.0 services use proprietary interfaces such as the Twitter API, the Facebook API or APIs from Google and/or they use lightweight interfaces and protocols, such as RSS or Atom. This works well in many cases, but in order to build more sophisticated services and service compositions there is a need for more sophisticated interfaces and concepts for interaction [1]. This can obviously be accomplished by using advanced proprietary APIs, as illustrated in [5], but from a wider perspective, common open standards are preferable. This is also one of the issues the study discussed in this paper is set out to examine. The next section describes the state of the art, followed by a brief discussion of some central concepts and ideas related to some previous work, followed by a description of the presented study and the experimental implementation of a Mashed-up PLE. Finally the results of the study are discussed in the light of the ongoing progress and previous research in the field.

A. State of the art

While LMS-like system are typically implemented by most educational institutions, the movement within the teaching community as well in the research community is towards adaptive and responsive learning environments, similar to PLEs, see e.g., [3][6][7][15][28]. However, while pedagogical concepts like responsive learning environments are attractive, the technology currently in use doesn’t support it very well. At the same time, education needs specialized services for dealing with pedagogical requirements, such as Personal Development Plans (PDP), digital portfolio, services for discovery and integration of digital learning resources, and so forth, which are resulting in several good and useful tools for learning, but they are not well integrated with the rest of the VLE [1][19][28]. These and similar issues are often addressed through different approaches to system integration, such as using proprietary APIs or more general integration by Web Service technology [1][4][5][14].

However, such approaches to building the Learning Infrastructure has turned out to be problematic for several reasons. Firstly, it becomes expensive to integrate "per system", using proprietary APIs. API integration also makes the systems hard coupled, which supresses flexibility [1][10]. Secondly, using (commonly SOAP-based) Web Service technology tend to become very complex as well as expensive, adding an cost, as well as technical, overhead [1][10][13], which is also illustrated in the VWE case discussed in section B. And thirdly, by mixing a monolithic concept, like the LMS with a modular service based approach some of the technical flexibility needed for dealing with some of the pedagogical requirements is lost [1][3][6]. In recent years there has been a general development on the Internet towards modularity and an alternative kind of loosely couple services driven by less complex and more web friendly service integration, such as using RESTFul APIs [21] and lightweight APIs and protocols, such RSS and Atom combined with widget and mashup technologies [16][17][20], which are described in detail in section C. This development stands out as exceedingly suitable for the next generation of learning environments, fulfilling the flexibility requirements for personal and responsive learning environments by providing a standardized framework for modularity and loose integration on the web that is now being studies by the research community in general and in an education context [25][28][30][31].

B. The Personal Learning Environment

Simply put, a *PLE* can be described as a learning environment where the learner is in focus as well as in control of the learning environment. However, the main objective of the PLE is to put the learner in control of his own learning rather than in control of the learning environment, even though these two are obviously related. Learning is regarded as a constant, ongoing process, as is the evolvement and change of the learning environment. The learning environment needs to be responsive and adapted to different contexts, needs and pedagogical requirements. These are qualities that are commonly emphasized, such as in [1], [3], [6] and [7], to give just a few examples. One of the ideas that are often emphasized in relation to PLEs is that personal "tools", such as blogs, twitter, etc., that are personal and used in other contexts can also be used as components of the PLE.

1) The Virtual Workspace Environment

The concept of a PLE is very similar (if not identical) to the idea underlying the *Virtual Workspace Environment* (VWE) that was first outlined in 1998, described in [2], even though the means to accomplish it were different. Simply put the VWE can be described as a component based VLE where users (i.e., teachers and students) can construct personal or shared learning spaces using a web browser.

In recent studies [1][5], it was shown that using modular approaches for the design and implementation of learning environments can address some of the LMS related issues, that were described in the previous section. A common modular taxonomy (*The VWE Learning Object Taxonomy*) for use with both VLEs and Digital Learning Resources (DLR) was presented in [2]. The taxonomy was compatible with the widely referenced *Learning Object Taxonomy* by Wiley [8] and demonstrated how the VLE and DLR could be implemented using a common modular, conceptual and architectural model that allowed for a common composition of both the VLE and learning content. Altogether this work resulted in two prototypes for composing and assembling modular VLEs; called the *Virtual Workspace Environment (VWE)*. The VWE was presented in [9], where the two different implementation approaches were compared. One using a JAVA RMI based approach and the other using a Web Service (SOAP) based approach. Both prototypes made it possible for teachers and/or learners to compose shared or personal learning environments by picking and choosing from a set of functional (software) components (called VWE tools). The VWE tools acted as building blocks providing the functionality for the learning environment. The ideas underlying the VWE were to a great extent inspired by the development of component-based software, as well as the fundamentals of Service Oriented Architectures (SOA), described in, e.g., [4][10][11][12].

A "proof of concept" was established, and by developing the prototypes using two different implementation approaches it was possible to isolate a couple of issues resulting from the taxonomy versus the model and the implementation approaches [9]. One of the problems that were identified was that, even though the use of standards was extensive (such as standards for Web Services, communication protocols etc.), the prototypes (and thereby the modular approach) only worked within the isolated context of the prototype environments and could not be generalized without new standards. This problem was mainly caused by a lack of standards supporting modularity for the creation of *Rich Internet Applications (RIA)* (see, e.g., [13]). In recent years, things have changed and standards have evolved and matured. Among the most interesting directions, from a modularity and RIA perspective, is the idea of Web Widgets and Mashups, see, e.g., [14][15][16][17]. The study presented in this paper starts out from the hypothesis that widget technology and widget mashups have the potential of overcoming many of the problems encountered during the VWE project [9], while still providing full support for the underlying ideas of modularity and the shift of central functionality and software from the desktop to the web, allowing for collaboration and social interaction with typical desktop functionality in ways that are only possible on the web. Furthermore, the creation of mashup learning environments can be adapted to different pedagogical scenarios and approaches in a dynamic and transparent way. Such transfer of functionality with its built-in potential has already been proven by services like Google Apps and other similar (web/cloud) services, see, e.g., [18] and [19]. However, the kind of rich functionality that is provided by such services needs to be put into context as an integrated part of the learning environment. Mashup Learning Environments (MUPPLE) are a step in this direction and it is also where the study presented in this paper and the WiMUPPLE project come in to play.

In retrospect, it can be said that the PLE concept is more Web 2.0 friendly and as such more flexible in terms of interpretation and implementation - with reference to choice and use (as well as "misuse") of technology, whereas the VWE concept provides a more explicit architecture model for the technology platform in relation to modularity and composition of mashup environments [9]. However, those features have also made VWE proprietary as only components that follow the VWE conceptual model and architecture can be used as building blocks. As a result, the VWE has also become too complex and dependent on VWE services and APIs as shown in Figure 1.

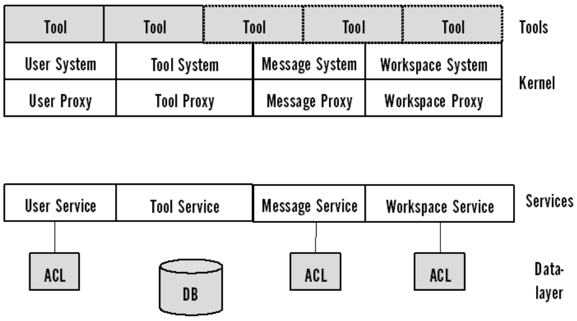


Figure 1. The figure shows an overview of the VWE architecture, the VWE Kernel and the VWE services used by tools to interact with the workspace.

In order for a component to work in the context of a VWE workspace, it needed to implement the VWE Service APIs, and all interactions with the workspace and other tools were via those server-side services. These were dependencies that severely limited the flexibility and usefulness of VWE from a Web 2.0 point of view.

For those reasons, one of the objectives of the WiMUPPLE project is to illustrate a third implementation strategy that addresses those problems and that makes the learning environment more generic, which is likely to be a characteristic needed in order for the concept of MUPPLEs to gain wider acceptance.

C. *Mashups and Widgets*

There are several (but similar) definitions of a mashup. A mashup is commonly defined as being a combination of different services on the web in a way that create a new composite application (or service) with added value. A widget-based mashup obviously uses widget technology and is currently typically constructed using a mashup environment such as Netwibes, iGoogle or our WiMUPPLE-environment [20]. A mashup can also be created by very simple means, using simple web tools that allow users to combine services on the web by matching and mixing information using lightweight interfaces such as RSS or Atom. However, in such cases it is mainly about mashing up information and not about mashing up functionality and services in a way that goes beyond the delivery and consumption of information. However, information mashups can be valuable in many cases, as part of a PLE.

Even so, if you are a developer or an experienced user you might want to use one of the more sophisticated approaches that are available for the development of web-based applications, or RIA as it is sometimes referred to.

The widget landscape is somewhat complex and can be roughly divided into three main categories: widgets for cellular phones (such as widgets for Android phones), desktop widgets (such as the widgets in OS X or gadgets in Windows) and finally web widgets, which are basically widgets that are distributed in the web browser [20]. *The widgets referred to in this paper are solely web widgets.* Even though these are three rather distinct categories there are several important similarities. One of the most significant similarities is that their implementations are based on web technology (or at least technologies that are commonly used on the web), such as html, JavaScript and XML (and AJAX). This is also an important property for sharing and reusing information and functionality since web technology relies on well-established standards. Besides the commonly used web standards there are widget-specific standards as well. However, widget standards are still rather untested and/or under development and there is still some way to go before it is possible to say that there are well established standards for widgets in the same sense as for the web. This means that there is always a trade off between the use of standards and proprietary widget technology when developing widgets that need sophisticated functionality.

The remainder of this paper presents and discusses a study that illustrates how widget technology can be applied to a modular concept, like the one previously described [2] and how a modular and web based learning environment can be implemented and assembled "on the fly" by learners and/or teachers. Both PLEs and LMS-like learning environments can be constructed in similar ways depending on the type of widgets, and supporting backend systems that are available. The same underlying SOA based server-side architecture that was used in the VWE project could in fact be used to support a client implementation using widgets, even though a REST based architectural model is preferred in order to avoid some of the complexity and limitations of the previous prototypes that were discussed above and in [9][21]. A RESTful approach also contributes to making integration with third

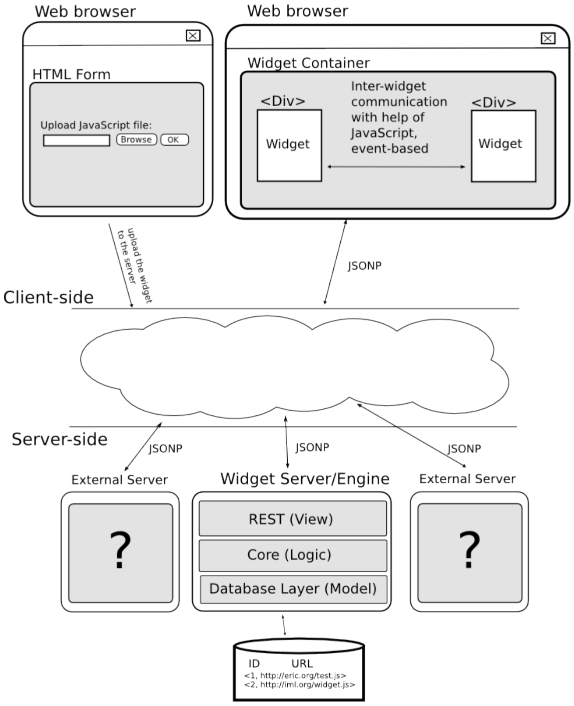


Figure 2. An overview of the WiMUPPLE-prototype architecture and its different parts with the Widget Container in the browser, interacting with the server layer via the REST API using JSONP.

party services easier and generally less complex and is more “web friendly”, even though not all problems can be solved in a RESTful way.

II. Objectives

The main objective of the research presented in this paper is to explore how widget-based mashups can be used as a basis for constructing a PLE or *Mashup Personal Learning Environments* (MUPPLE) as PLEs are referred to when implemented as mashups. The mashup approach can be compared to the two approaches used in the VWE project that was briefly described above.

In addition, the widget based MUPPLE approach is applied to a similar modular concept that was presented by Paulsson and Berglund in [9], where it was illustrated how the modular concept of Learning Objects can be extended to also become a modular concept for the whole VLE (i.e., a PLE or an LMS) by adding some basic software architectural rules and principles that conjures a number of essential properties to the otherwise content centred concept of Learning Objects, see [22][2][1].

Another objective is to illustrate that the concept of a modular framework, such as the VWE Learning Object Taxonomy, can be applied using more Web 2.0 friendly and generic approaches. Therefore the work presented in this paper will be discussed and compared to the work presented in [2], where the VWE Learning Object Taxonomy was introduced and in [1][9], where the two VWE prototypes were discussed (also discussed above) in relation to pedagogical requirements and learning theories.

III. Methodology

Besides surveying the literature in the field, this study is based on an experimental approach where a prototype was developed and tested. It should be emphasized that even though this study addresses issues and requirements that emanate from a pedagogical standpoint, i.e., creating conditions for pedagogical adaptability and responsiveness in technology, the objective is not to evaluate the pedagogical implications at this point. The purpose is instead, which is also discussed above, to evaluate how the concept of a more generic and web friendly approach, using widgets and mashups, can be utilized from a technological standpoint to build MUPPLE, in comparison to earlier less generic implementations, such as the VWE. And furthermore - how to do this by applying existing concepts. However, in the discussion section of the paper the results are also discussed in relation to some pedagogical issues and implications based on experience from other studies, in order to better illustrate how modularity, technology implementations and pedagogical issues are linked.

1) *Technology settings*

An important starting point was to avoid developing everything from scratch. There are a multitude of ongoing development and project addressing widgets and mashups and whenever it has been possible existing work has been used.

The prototype architecture follows common design paradigms and patterns, illustrated by Figure 2, which also illustrates how widgets are handled on the client using a widget container that renders the widgets. The inner working of the widget container is illustrated in Figure 3 and described in more detail below. Even though Figure 2 illustrates a schematic architecture using a web browser as the client, the client could in fact be any other widget platform, such as a handheld device or dashboard widget. The widgets used for the purpose of the prototype are described from the point of view of being used in the context of a VLE, but in theory most of the widgets could be used in other contexts as well as they are often generic functional components that have been contextualized by the mashup and the pedagogical context.

JavaScript Object Notation (JSON) [23] is used for communication and data interchange between widgets and servers. As illustrated by Figure 2 and Figure 3, widgets that run on the WiMUPPLE platform can interact with a widget server, a widget engine or any other external server using JASON (or JSONP for managing cross domain interaction). This creates a flexibility that goes beyond the “local” ICT infrastructure and makes it possible for widgets to potentially interact with any servers that are of interest, acting as lightweight clients for other systems that may be relevant in the context of a learning environment. This differs from the previous VWE implementations in that it provides a transparent and generic infrastructure rather than a proprietary and platform dependent API.

This creates flexibility in terms of making the learning environment adaptable to different and chancing requirements. Furthermore, it makes the learning environment independent of a specific LMS vendor to implement certain functionality. It has proven to be quite straightforward to develop simple widgets that can act as clients to different legacy (as well as to other) systems.

Flexibility, in terms of being adaptable and distributed, is an essential property of a modular environment since it allows for the learning environment to be distributed (service-wise) over the Internet and at the same time it makes it possible to personalize and adapt the learning environment at the service level for group preferences as well as for personal preferences. It also creates the characteristics needed for responsive VLEs. These are important differences compared to the concept of an LMS, which has a centralized approach with clear system borders limiting the ability to interact with the surrounding world to the interactions that are countered by the LMS vendor or (in some cases) plug-ins and suchlike developed by third-parties, This also means that the functionality is limited to what is supported by the LMS, while functionality can be added and removed dynamically in the mashup PLE.

B. *The Widget Container*

The client hosts the widgets within the *widget container* (see Figure 2), which is loaded into the browser and rendered. Each widget has the possibility to communicate and interact with external servers as well as “internal” widget specific servers that are specifically developed to serve the widget. The widget container can actually be compared to the “kernel” in the VWE implementation. However, the kernel was implemented as a Java Applet, while the widget container relies on the JavaScript capabilities of the web browser and the standards associated with widgets and is therefore a more generic solution. Figure 1 illustrates the VWE kernel implementation, while Figure 2 illustrates the role of the Widget Container.

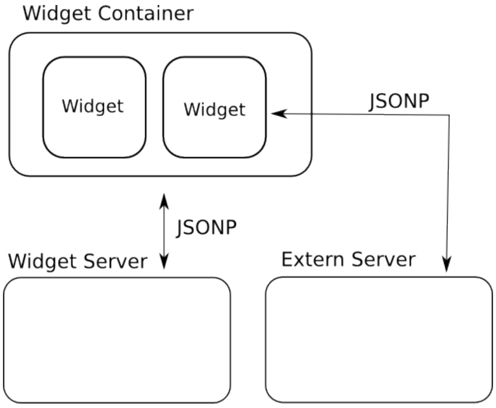


Figure 3. Illustrates the design and inner works of the Widget Container and how widgets interact with the widget server and/or external servers using JSONP.

Technically, the widgets used by the system consist of JavaScript that is loaded into the widget container where they are rendered and executed. The Widget Server keeps track of what widgets are available and the Widget Container communicates with the Widget Server using JSONP and the predefined RESTful API. Thanks to the Widget Container, it is possible to move the widgets around in the browser's workspace. Hence, the Widget Container also serves as the "glue" that holds the browser representation of the learning environment together and creates the feeling of an integrated environment in the same sense as the LMS. There is however an essential difference in the philosophy and approach underlying the integration. While the LMS relies on a strong, silo-like strategy for integration, the MUPPLE relies on loose integration of freestanding services and components.

C. *The Widget Server and the Widgets*

As previously mentioned, widgets are basically JavaScript uploaded to the Widget Container via the Widget Server. Besides the communication with the Widget Server, widgets can communicate with other external servers using JSONP. In the case of WiMUPPLE a choice was made to use *Yahoo Querying Language* (YQL) [24] for the implementation of the widget server in order to avoid unnecessary in-house development. However, it is fully possible to use other approaches as well, such as Google or other servers that are widget specific, with similar results. This is actually not a big issue and is illustrated by Figure 2. Besides YQL, the WiMUPPLE Widget Server was built using the Python-based Django framework and a traditional MVC pattern.

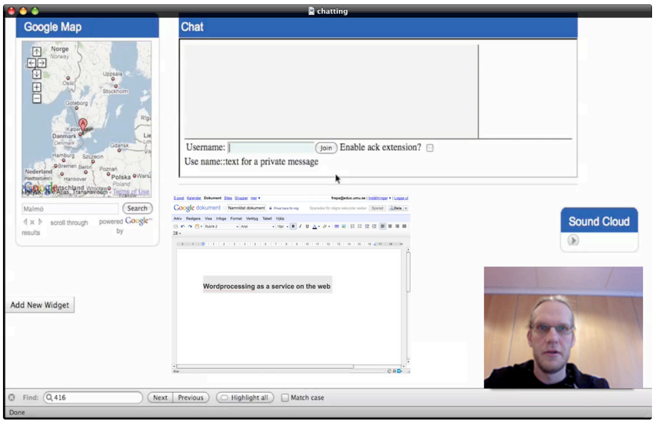


Figure 4. Screenshot of a learning environment created using WiMUPPLE with a number of widgets for different purposes.

IV. Results and Discussion

The experimental implementation and the resulting prototype show that it is quite possible to implement a modular VLE using widgets and mashup technology. Or in the WiMUPPLE case, a framework for composing and administrating mashup learning environments in a way that can be managed by teachers and students (shown by Figure 4) and in which functionality can easily be added and removed. With the right set of widgets, a complete LMS could theoretically be built using the WiMUPPLE, even though an LMS is probably not what is wanted or needed.

The WiMUPPLE implementation makes it quite clear that it is less complex using a widget approach compared to the Java RMI and/or SOAP approaches used in the VWE-project [5]. However, the widget mashup approach is in general less powerful in terms of building sophisticated functionality. One of the main issues in this respect is the internal communication, i.e., inter-widgets communication. In an LMS everything is closely integrated, which is also what causes the main problem with the LMS concept, but at the same time it is a strength in terms of inter-system communication. All parts within the system can easily be made aware of all other parts in the system. In a mashup everything is loosely coupled and different widgets are normally completely self-standing and self-contained and not "aware" of the context in which they are used. This makes it harder to maintain the feeling of a well-integrated learning environment. The VWE implementation had similar problems that were solved by implementing a "Message Service" (see Figure 1) that managed the interaction between components and different parts of the learning environment, including other components (tools). The drawback with this solution, besides being proprietary, was that all the components became dependent on a common server side infrastructure in order to function in the context of an integrated learning environment. When working with widget-based mashups such solutions become a problem, as we want to be able to use any kind of widgets that follow the widget standards, i.e., not depending on a common server side infrastructure. An alternative solution would therefore be to make the widgets aware of each other within the web browser and allow widgets to interact and communicate with each other directly. This is technically possible and Sire et al. have described an example of such interaction in [25], where they discuss the implementation of drag and drop between widgets in the browser. There is currently no standardized or obvious way of implementing direct widget interaction and it will demand some tweaking to work. However, this is one of the issues that are likely to be solved by html5.

There has been some tweaking in order to get everything to work as expected, which was mainly caused by the immature nature of the widget technology concept compared to the maturity of Java RMI and SOAP. However, it is highly likely that the adoption of html5 will solve many of the problems and issues encountered here as well.

The fact that the principle of modular VLEs is realistic was shown in [1][9] it was discussed how modularity contributes to creating important pedagogical advantages related to flexibility and adaptability, which are qualities needed to create learning environments that are responsive and adaptive to users needs and to changing pedagogical requirements. The WiMUPPLE add to those characteristics by using generic and web friendly technology that open up for a much wider range of components to choose from.

Furthermore, modular learning environments are better adapted to suit different learning theories and pedagogical approaches as well as to changing pedagogical scenarios. Such features are beneficial, even essential, in many learning scenarios, especially when working with pedagogical methods and approaches like Learning, where it is hard (if not impossible) to foresee the learning path from start to finish beforehand – and thereby also to foresee the needs of the learning environment. These are also the main reasons why it is important to continue the research and development of modular concepts for VLEs - like the WiMuPPLE. Taken as a whole, the project also illustrates potential business cases where market competition is opened up for smaller actors to compete with LMS vendors by providing small and specialized components acting as building blocks in a mashup learning environment.

It has already been shown that modular learning environments hold an interesting pedagogical potential. In [1], it was illustrated that there is a correlation between modular environments and adaptability and responsiveness and that such features create pedagogical flexibility. The experimental study presented in this article shows that, not only is it possible to build modular learning environments, but it can be done using web based standard technology that bears the potential of almost endless flexibility in terms of access to functional components – in this case widgets. In the long run, this means that generic components (i.e., widgets) can be integrated as a part of a modular VLE without the need of adding learning specific code or support for certain APIs, even though such APIs may be beneficial in many cases, something that is discussed below. Even though the experiments showed that this can be technically accomplished (even if the technology is still somewhat immature) there is still a need for a better “glue” to tie mashup learning environments together and to create pedagogical context.

A. 3.2 Future work and developments

There are some very intriguing progresses around the corner that are likely to benefit the development of mashup learning environments. On the one hand there is the general development, such as the gradual evolvement of html5 and standards for widgets and mashups. On the other hand there are developments within the field of learning technology standards that, at least on paper, look very promising from a modular learning environment perspective. Among the most interesting developments are the new specifications from IMS: IMS Common Cartridge (IMS CC) [26] and especially the IMS Learning Tools Interoperability (IMS LTI) [27]. We are currently in the process of examining whether IMS CC can be used as a packaging format for our widget-based MUPPLE and furthermore, if IMS LTI can be used as a standard for widget communication and interactions within a widget-based VLE. Severance et al. have already described some experiments in [28] where IMS LTI was tested in a mashup environment and the results seem promising and could be taken even further in the WiMUPPLE environment.

Another direction, that has already started, is the integration of the *Spider* and the WiMUPPLE environment. The Spider is a national search service for digital learning resources that connects a number of repositories, using either metadata harvesting or search federation, in a way that makes it possible to search for learning resources from several sources from a single point [29]. The idea is to use the Spider to search for Widgets and learning content that can be included in a mashup learning environment and then use IMS CC to package them into a “package” that, when unwrapped, constitutes a mashup learning environment. In conjunction to this it seems reasonable to start discussing digital learning resources from a broader perspective – not just being about learning content, but also functional components, such as widgets.

In parallel with the developments described in the previous section, another project will start where some pedagogical experiments will be carried out using the WiMUPPLE environment, where the idea of mashup learning environments will be tested in real pedagogical situations with students and teachers.

References

[1] F. Paulsson. Modularization of the Learning Architecture: Supporting Learning Theories by Learning Technologies. Royal Institute for Technology (KTH): Stockholm, 2008. 121.

[2] F. Paulsson and A. Naeve. "Virtual Workspace Environment (VWE): A Taxonomy and Service Oriented Architecture Framework for Modularized Virtual Learning Environments - Applying the Learning Object Concept to the VLE". International Journal on E-Learning, 2006, vol. 5, no. 1, pp. 45-57.

[3] G. Atwell. Personal Learning Environments - the future of eLearning? [online]. 2007, [Retrieved: April, 2012]. Available from World Wide Web: www.elearningeuropa.info/files/media/media11561.pdf

[4] T. Erl. SOA Principles of Service Design. 1 ed. Prentice Hall: Boston, MA, 2007.

[5] F. Paulsson. "A Service Oriented Architecture-framework for modularized Virtual Learning Environments," In A. Mendes-Vilas, A. Solano Martin, J. Mesa Gonzáles, and J.A. Mesa Gonzáles, Current Developments in Technology-Assisted Education. FORMATEX, 2006, pp. 21-62.

[6] S. Wilson, O. Liber, M. Johnson, P. Beauvoir, P. Sharples, and C. Milligan. "Personal Learning Environments: Challenging the dominant design of educational systems," First European Conference on Technology Enhanced Learning (ECTEL), 2006,

[7] D. Jones. "PLES: FRAMING ONE FUTURE FOR LIFELONG LEARNING, E-LEARNING AND UNIVERSITIES". Lifelong Learning Conference, 2008,

[8] D.A. Wiley. "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy". In D.A. Wiley, The Instructional Use of Learning Objects. Bloominton: Agency for Instructional Technology and Association for Educational Communications & Technology, 2002, pp. 3-23.

[9] F. Paulsson and M. Berglund. "Suggesting a SOA-framework for modular virtual learning environments: comparing two implementation approaches," International Journal of Web-Based Learning and Teaching Technologies, 2008, vol. 3, no. 1, pp. 43-57.

[10] P. Brereton and D. Budgen. "Component-based systems: a classification of issues. Computer", 2000, vol. 33, no. 11, pp. 54-62.

[11] O. Nierstrasz and L. Dami. "Component-Oriented Software Technology," In O. Nierstrasz and D. Tsichritzis, Object-Oriented Software Composition. Hertfordshire: Prentice Hall International (UK) Ltd, 1995, pp. 3-28.

[12] C. Szyperski. Component Software - Beyond Object-Oriented Programming. Translated by C. Szyperski; 2 ed. ACM Press: New York, 2002. 229.

[13] J. Preciado, S. Comai, and C. Sánchez-Figueroa. "Necessity of methodologies to model Rich Internet Applications,". In the proceedings of the Seventh IEEE International Symposium on Web Site Evolution (WSE’05), 2005.

[14] M. Ogrinz. Mashup Patterns Designs and Examples for the Modern Enterprise. Addison Wesley: 2009.

[15] S. Sire and A. Vagner. "Increasing Widgets Interoperability at the Portal Level," The First International Workshop on Mashup Personal Learning Environments (MUPPLE-2008), 2008.

[16] J. Yu, B. Benatallah, F. Casati, and F. Daniel. "Understanding Mashup Development," IEEE Internet Computing, 2008, vol. 12, no. 5, pp. 44-52.

[17] J. Wong and J. Hong. "What do we ”mashup” when we make mashups?" In the proceedings of the 4th international workshop on End-user software engineering, 2008.

[18] D.R. Herrick. "Google this!: using Google apps for collaboration and productivity," SIGUCCS '09 Proceedings of the 37th annual ACM SIGUCCS fall conference, 2009, pp. 55-64.

[19] N. Sultan. "Cloud computing for education: A new dawn?" International Journal of Information Management, 2010, vol. 30, no. 2, pp. 109-116.

[20] V. Hoyer and M. Fischer. "Market Overview of Enterprise Mashup Tools," Service-Oriented Computing, ICSOC 2008. In Lecture Notes in Computer Science, 2008, vol. 5364, pp. 708-721.

[21] Fielding, R. T. 2000. Architectural Styles and the Design of Network-based Software Architectures. Information and Computer Science. Doctor of philosophy, 76-106.

[22] F. Paulsson and A. Naeve. "Establishing technical quality criteria for Learning Objects,". in the proceedings of eChallenges 2006, 2006, vol. 2, pp. 451-462.

[23] D. Crockford. The application/json Media Type for JavaScript Object Notation (JSON) [online]. 2006, [Retrieved: April, 2012] Available from World Wide Web: https://tools.ietf.org/html/rfc4627

[24] Yahoo! Query Language [online]. 2011, [Retrieved: March, 2012] Available from World Wide Web: http://developer.yahoo.com/yql/

[25] S. Sire, M. Paquier, A. Vagner, and J. Bogaerts. "A messaging API for inter-widgets communication,". In the proceedings of the 18th international conference on World Wide Web, 2009,

[26] IMS Common Cartridge Specification [online]. 2008, [Retrieved: April, 2012]. Available from World Wide Web: http://www.imsglobal.org/cc/index.html

[27] IMS GLC Learning Tools Interoperability Basic LTI Implementation Guide. 2010, vol. Version 1.0 Final Specification.

[28] C. Severance, T. Hanss, and J. Hardin. "IMS Learning Tools Interoperability: Enabling a Mash-up Approach to Teaching and Learning Tools," Techn, Inst, Cognition and Learning, 2010, vol. 7, pp. 245-262.

[29] F. Paulsson. "Connecting learning object repositories - strategies, technologies and issues". In the proceedings of the Fourth International Conference on Internet and Web Applications and Services (ICIW09), 2009 pp. 583-589.

[30] Weber, N., Nelkner, T., Schoefegger, K., and Lindstaedt, S. N. 2010. SIMPLE - a social interactive mashup PLE. the Third International Workshop on Mashup Personal Learning Environments (MUPPLE09), in conjunction with the 5th European Conference on Technology Enhanced Learning (EC-TEL2010).

[31] Wheeler, S. 2009. Learning Space Mashups: Combining Web 2.0 Tools to Create Collaborative and Reflective Learning Spaces. Future Internet. 1, 1, 3–13. DOI=10.3390/fi1010003.

SecureRoutingDHT: A Protocol for Reliable Routing in P2P DHT-based Systems

Ricardo Villanueva*†, María-del-Pilar Villamil*
*Department of Computer and Systems Engineering
University of Los Andes, Bogotá, Colombia
Email: rvillanueva@egresados.uniandes.edu.co,mavillam@uniandes.edu.co
†University Simon Bolivar, Barranquilla, Colombia
Email: rvillanueva1@unisimonbolivar.edu.co

***Abstract*—Secure routing in P2P distributed hash table based systems has been an open subject for several years due to the importance of the routing protocol in these systems. Providing security at the routing level is a hard task because of the open nature of these systems. This article presents a protocol for reliable routing in P2P DHT-based systems, which mitigates routing attacks. It makes use of a quorum topology and a reputation system to provide security at the routing level. It is shown, theoretically and by simulations, that proposed protocol keeps stable the number of involved messages in the forwarding process, as well as it tolerates up to 30% of malicious nodes.**

***Keywords*-P2P DHT-based systems; security, threats; routing; quorum; reputation; Bayesian-systems.**

I. Introduction

Distributed hash table based systems are a special class of distributed system with interesting properties such as scalability, decentralization and self-organization. On top of these systems have been built a plenty applications such as distributed storage, application layer multicast, and so forth [5]. Nevertheless, on building of these applications has not been considered security as a main quality attribute, for that exists several threats to be taken into account.

Providing security to these systems is rather challenging due to DHT inherent features. According to Sit and Morris [28], threats against these systems comes from anywhere. In fact, they identified several attacks, and further classified them into routing, storage/retrieval and miscellaneous attacks. Particularly those threats against routing mechanism are extremely important, since they could compromise the proper functioning of the whole system.

The routing process is composed of two main sub-processes: routing table maintenance and message forwarding. Therefore, a malicious peer could misroute or drop messages along the path -*incorrect lookup*-, attempt to corrupt routing table entries of other nodes - *eclipse* attack [27][6])-, fool any peer through joining process in order to induced it into an incorrect network -*overlay partition*-, send unused messages or frequently joining/leaving the overlay network.

Although there are several works that have addressed this problem [33], these works are isolated, namely, only focusing on a specific system or attack; they even do not consider performance issues as number of messages. This paper presents a protocol that extends the underlying DHT to a redundant topology and makes use of one reputation mechanism in order to harden the DHT, but mantaining the number of sent messages stable and being easily coupled to other mechanisms.

This paper is organized as follows: Section II presents models and assumptions, which are used throughout all this paper. Section III presents strategies proposed to mitigate the impact of the *routing* attack. Section IV presents the reliable routing protocol SecureRoutingDHT. Section V presents the theoretical and practical (through simulations) analysis. Finally, Section VI concludes and gives some perspectives about future works.

II. Assumptions and definitions

Each DHT system is defined over an identifier space K, where peers and resources are mapped into. A closeness metric ρ is used for matching resources to peers. Commonly, this is achieved by using a proper hash function h, defined from the peers/resources set to K [5].

Furthermore, each node p within the system has at least two different types of links to other nodes, namely, p maintains links to specific close and distant nodes. Those links form the so-called *routing table*, which is used in the forwarding process. This process uses a greedy algorithm that has been implemented in three different ways: recursive, iterative and trace [33].

In *recursive* routing, a initiator p requests for a resource and *consequently this request* is forwarded by each intermediate peer to a next one independently. Whenever this request has reached to the responsible node, the *reply* is sent directly to the initiator or forwarded back by peers on the reverse path. Unlike previous, in*iterative* routing each intermediate peer sends contact information of next peer back to the initiator, hence p will be able to contact directly to the next peer. As a consequence, *p* is able to detect misbehaviour peers through techniques based on structure of DHTs, nonetheless the latency of the forwarding process is augmented.

Finally, *tracer* routing is a combination from both previously introduced techniques [35]. In this mechanism, each intermediate node must forward the message to next peer but also sends contact information of next peer to the initiator. Therefore, *p* knows about the entire process but routing latency is not affected.

In connection with the security model to be considered, it is supposed that there is a mechanism that randomly assigns a $nodeid$ to each new peer. In fact, this can be accomplished by coupling our proposal to other ones whose goal is to mitigate the *Sybil Attack* [1][17]. As a consequence, only a fraction of malicious peers exist during a period of time, as well as peers are uniformly distributed over the DHT. Moreover, a malicious peer can discard, generate or incorrectly forward any message. This model is widely known as *random fault model* [4], where a peer is malicious with probability at most f.

III. Secure Routing in DHT-based Systems

There are several proposals that try to mitigate *routing attacks*. In previous work [33] , we classified these strategies based on how attacks are addressed, namely, we identified three styles of solutions: those based on message redundancy, those based on malicious node detection and those trying to avoid malicious nodes by computing trust profiles of other peers.

As far as redundancy-based strategies are concerned, the requester sends several messages in order to increase the probability of reaching the responsible peers. In this style of solution, two approaches were identified: multi-path routing, where the requester peer sends a message among its neighbours, hence it is being forwarded to the responsible peers through multiple paths [4][11][15][23]. On the other hand, in wide-path routing strategies, peers are re-arranged in groups (*quorums*), hence the initiator peer broadcasts the request to each peer within its group, afterwards the message is broadcasted at same way by other peers until it reaches the destination quorum [19][24][36].

Concerning malicious node detection techniques, the sender detects a malicious node by verifying whether an invariant of the system is fulfilled - one of the most used invariant is that each hop is closer to the target during *lookup* process-. Otherwise, the sender requests to a previous considered-good node for another peer, in order to continue with the search [20][21][29][34].

Incidentally, strategies based on trust profile attempt to measure, under a well defined mechanism, which peers are the more suitable in the forwarding process. These mechanisms have been implemented using reputation systems and social networks. The former allows each peer to construct the profile of other peers using historical data and recommendations [9][21][24][25][26].

Conversely, those using social network build the trust profile of peers based on features of the social network graph. For instance, Sprout [18] relies on the fact that friends are expected to have a more reliable behavior than other ones. On the other hand, the technique introduced by Danezis et al. [7], is based on sending requests to peers which have appeared a few times in social graph paths, therefore requests are balaced over the system.

IV. SecureRoutingDHT

This section introduces SecureRoutingDHT, a protocol to provide reliability in the lookup process. In essence, this section presents decisions that were taken into account for constructing this protocol.

A. Routing Protocol Construction

The routing protocol is defined over a redundant structured topology that is organized into several groups of peers called *quorums*. These groups are connected among them and are constructed by augmenting the routing table. A *quorum* provides flexibility and diversity for selecting a peer during the lookup process, as well as storing multiples replicas of an object and cooperating among peers.

Each peer p in a redundant topology, maintains three levels for routing information:

1) Peers within its *quorum*: the peer p has links to all peers in its quorum.
2) Peers in other *quorums*: For each contact peer, q, of p, it maintains links to all peers within the quorum of q.
3) Backpointers: p maintains pointers to the peers pointing to it.

The aforementioned construction suggests that the overlay structure is strongly connected (redundant). Hence, some of the attacks previously introduced are hardly to lunch. In fact, with this structure, each node can verify their links so as to detect lunched routing poisoning and unsolicited message attacks.

1) Protocol: SecureRoutingDHT makes usage of the *recursive* style routing; but, during the process, each peer is provisioned of a selection function (Reputation mechanism) that chooses the most reliable peer within the next *quorum*, to send the request. Finally, at penultimate node, the request is broadcasted to a subset of peers within the last *quorum*, thus resistance to storage and retrieval attacks is provided. Figure 1 illustrates, in a general fashion, how the routing process is performed by SecureRoutingDHT.

Let Q_k be a quorum at $k-th$ step of the routing process, R_{qp} be the reputation of peer q maintained by p and h be the average number of steps to reach the

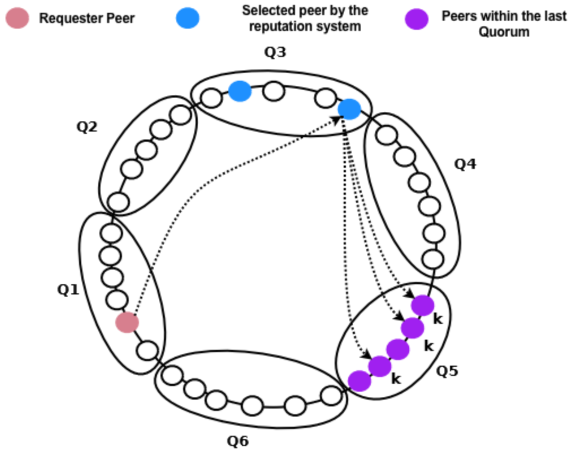

Requester Peer
Selected peer by the reputation system
Peers within the last Quorum
Q3
Q2
Q4
Q1
k
k
k
Q5
Q6

Figure 1. Routing Process

target. Now, let us suppose a peer $p \in Q_1$ (initiator) requests for key k, so the protocol works as follows:

- p constructs a set of peers, S , whose reputation value exceeds a threshold u (configuration parameter). Formally, $S = \{q \in Q_2 | R_{qp} \geq u\}$ and for completeness $S = Q_2$ if $S = \oslash$. Now, peer p randomly chooses a peer q_2 from S and sends the request. This random selection allows feeding the reputation mechanism.
- Let $p = q_2$ and repeat step 1 until reaching *quorum* Q_{h-1}.
- Finally, q_{h-1} sends the request to each node within a subset $D \subseteq Q_h$, which is responsible for storing k.

2) Routing Protocol Maintenance: Each overlay network needs a process to keep up to date the organization of the peers within it. This process is time consuming since it is performed frequently, but it is even more consuming in quorum based overlay networks because of its higher number of links among peers. However, there is a trade-off between security and performance.

Whenever a peer p is joining to the network, it contacts another peer q by sending its corresponding membership *token*. At that point, q validates the *token* sent by p. Afterwards q forwards a *join* message with identifier id_p. As soon as q receives the responsible *quorum*, it sends back to p. At that stage, p is able to notify to each neighbour, allowing them to update their routing information. Finally, p performs a similar process by sending multiple queries, which depends on the underlying DHT, in order to build of other *quorums*. Moreover, for each formed *quorum*, p notifies them, so as that they can update their *backpointer* information.

B. Reputation Mechanism

The proposed reputation mechanism was designed by taking into account three components suggested by Hoffman et al. [12]. These components are: (1) Formulation, (2) Calculation, and (3) Dissemination. The first one defines the reputation metric foundations and the information sources. In turn, the second one describes the formulae to compute the reputation of a peer, and finally, the third one defines the interaction mechanisms among peers.

1) ***Formulation****:* Since the routing process is *recursive*, a peer is only able to compute ratings in accordance with the success or failure of sent messages. Thus the reputation-updating process is realized by asking recommendations or through own interactions. Accordingly a 4-tuple (E_e, E_f, R_e, R_f) is defined, where variables E_e and E_f are two events, reprsenting if a message is forwarded successfully or not respectively. In turn, R_f and R_e are events representing whether a recommendation is considered as biased or non-biased respectively. As it can be noted, a peer only assigns reputation values to peers within its routing table.

2) ***Calculation****:* There are several ways to compute the reputation of a peer, namely, average ratings, trust models or Bayesian models and so forth [13]. On the one hand, models based on simple average are not appropriate since they do not allow representing the context adequately. On the other hand, trust models and Bayesian systems, which have been extensively studied and proved as equivalent, are more adequate because of their properties such as context evaluation, easy computation, extensibility in terms of number of variables and aging [14]. Moreover, Bayesian reputation systems are those based on the Dirichlet function, for that allowing the definition of several variables [14].

Let $X = \{X_1, X_2, \ldots, X_k\}$ be the set of k random variables, which represent the events of the observations and p_i be the probability function for X_i which satisfies $\sum_{i=1}^{k} p_i = 1$.

The computation Dirichlet function is not practical; as a consequence this value is calculated as [14]:

$$\wp = \Sigma_{j=1}^{k} \tau_j \vec{S}(X_i) \qquad (1)$$

where $\vec{S}(X_i) = E(\vec{p}(X_i)|\vec{r}, \vec{a}) = \frac{\vec{r}(X_i)+W\cdot\vec{a}(X_i)}{W+\Sigma_{j=1}^{k}\vec{r}(X_j)}$ is the expected value of X_i and values $1 \leq \tau_j \leq k$ are weights.

Note that, if $\tau_j = 1$ for all $1 \leq j \leq k$, $\wp$ will be equal one. In addtion, $\vec{a}$ is the base rate vector over the state space and W is a weight, which is typically set to 2, but W could be chosen higher in order to reduce the influence of new evidence over the base rate [14].

Observations are accumulated as a vector $\vec{R} = (R_1, R_2, \cdots, R_i, \cdots, R_k)$ by each peer. If an event which affects to variable X_i is detected, $\vec{R}$ will be updated by performing $\vec{R} = \vec{R} + I_i$, where I_i is the identity vector.

For aging observations, let $M_{y,t}$ be the set of peers that collect observations during a interval time *t* for an agent y, $\vec{r_y^x}$ be the vector of observations collected by x for y in the same interval. Now let $\vec{r_{y,t}}$ be the set of the total observations in the interval t for agent y, hence $\vec{r_{y,t}} = \Sigma_{x \in M_{y,t}} \vec{r_y^x}$. Furthermore, vector $\vec{R}$ can be updated by computing $\vec{R_{y,t}} = \lambda \vec{R_{y,t-1}} + \vec{r_{y,t}}$, where $0 \leq \lambda \leq 1$. As it can be noted, a higher value of λ gives more priority to historical data.

Finally, the reputation of a peer is calculated as follows

$$\wp = \tau_{E_e}\vec{S}(E_e) + \tau_{R_e}\vec{S}(R_e) + \tau_{E_f}\vec{S}(E_f) + \tau_{R_f}\vec{S}(R_f) \qquad (2)$$

where $\tau_{E_e} = 1$, $\tau_{R_e} = 0.5$ and $\tau_{E_f} = \tau_{R_f} = 0$. These assignments give a higher priority to successful messages, because they are performed more frequently.

3) ***Dissemination****:* There are two sources of information: direct and indirect. The former encloses the interactions which a peer has with other peers, and the latter comprises the provided recommendations from a peer.

The recommendation process builds a set of peers built from known *quorums* and asks for a recommendation to each peer within it of another peer. Each provided recommendation is sent back by using the *Piggyback* protocol. As soon as recommendations are received, those are classified as biased or non-biased by performing the following classification algorithm:

Suppose that p asks for recommendations for a peer r to a group of peers C_r. At that point, there is expected that each peer c within C_r sends to p the corresponding recommendation as a vector $\vec{R^c_{r,t}}$. As soon as a defined time has elapsed, p computes the local reputation value of r, $\wp_r^p$, as well as $\wp_r^c$ for each received recommendation from $c \in C_r$. These values are computed with $\vec{R^c_{r,t}}$ and the local base vector $\vec{a}$. Afterwards, p computes the following

$$\sigma = \sqrt{\frac{\Sigma_{c \in C_r}(\wp_r^p - \wp_r^c)^2}{|C_r|}} \qquad (3)$$

Now, let us consider the interval $I = [\wp_r^p - k \cdot \sigma, \wp_r^p + k \cdot \sigma]$, where k is a positive constant that generally is setted to 1. The classification method arranges each received-recommendation $\wp_r^c$, as biased or non-biased, if $\wp_r^c$ is within the interval or not respectively. For those peers which sent a considered-biased recommendation, the corresponding variable R_f is incremented by 1, otherwise the corresponding variable R_e is incremented by 1.

Furthermore, a new set of recommendations, E, is formed with each one of received recommendation considered as non-biased- these recommendations are represented as a vector-. From the set E, p only chooses a few recommendations in order to avoid that colluding peers try to overstimate/understimate the reputation value of another peer; and updates the corresponding reputation value by computing $\vec{R^p_{r,t}} = \vec{R^p_{r,t}} + \Sigma_{c \in H_r}(\wp_c^p \cdot \vec{R^c_{r,t}})$.

Besides of mentioned components, it is important to define a mechanism to reduce the impact of churn to the reputation system. In fact, a peer can take of advantage of the joining/leaving process to gain a new reputation value [25]. Therefore, a mechanism that alleviates this threat must be implemented.

A possible solution is to use the same system to store these values, even though this would imply an increment of the number of messages, as well as adressing new concerns - those related to data availability, integrity, privacy and access controls [22]. Therefore, this sort of solution is not appropiate.

As a consequence, another strategy is implemented which takes advantage of the fact that several Sybil attack solutions assign a fixed identifier to each peer [1][4][16][17]. Following this, it is likely that the set of backpointers of the joining peer would be the same, consequently a local cache is proposed in order to store calculated reputation values of the off-line neighbours.

Since cache size is finite, the implemented replacement policy only maintains reputation values of those peers which are likely to rejoin to the system (LRU-based). Each peer is assigned a default estimated off-line period, called PER_0 at first time. In this way, whenever a peer leaves/joins the system, its backpointers peers calculate a off-line period PF and update the correspoding PER by computing PER_{i+1}=$PER_i \times \alpha$+$PF \times \beta$, where α, β are weights which tipically are set to 0.2 and 0.8, respectively [3].

V. Evaluation

On this section is presented an analysis of our protocol regarding the number of messages during its operations, as well as the probability of success whenever messages are forwarded.

A. Theoretical Analysis

Theoretical analysis is presented regarding number of involved messages in the forwarding and maintanance proccess, as well as the expected probability for that a message reaches to responsible peers.

1) Number of messages: Suppose that $q_1 \in Q_1$ is searching the responsible nodes of one resource with id k. Let $Q_1, Q_2, \cdots, Q_h$ involved *quorums* during the routing process. Note that h depends on the underlying P2P DHT-based system. Moreover, let $D \subseteq Q_h$ be the set of peers storing key k. Hence, the expected number of messages to reach D is equal to $h-1+|D|$. Particularly, If Chord is the underlying system, there holds that $h = O(\log_2 n)$.

TABLE I
EXPECTED NUMBER OF MESSAGES

Strategy	Number of expected messages	Error
Chord [30]	$\log_2 n$	0
Replica [11]	$2^{\lvert D\rvert-1}\log_2 n$	$\log_2 n \cdot (2^{\lvert D\rvert-1}-1)$
S-Chord [10]	$\lvert D\rvert^2 \log_2 n$	$\log_2 n(\lvert D\rvert^2-1)$
QuorumP1 [36]	$2 \cdot \lvert D\rvert \log_2 n$	$\log_2 n(2 \cdot \lvert D\rvert - 1)$
QuorumP2 [36]	$\frac{(\log_2 n-1)}{1-f} + 2 \cdot \lvert D\rvert$	$\frac{f \cdot \log_2 n-1}{1-f} + 2 \cdot \lvert D\rvert$
Secure Routing DHT	$\log_2 n - 1 + \lvert D\rvert$	$\lvert D\rvert - 1$

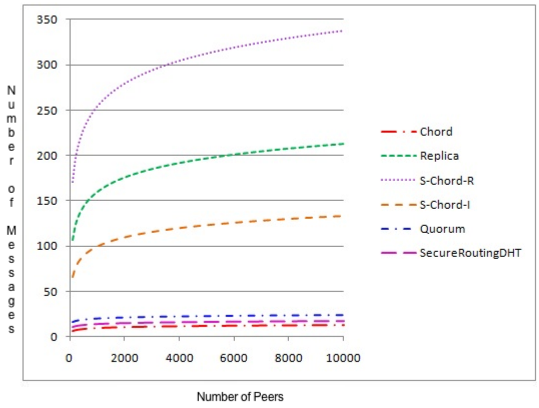


Figure 2. Expected number of messages

Table I presents formulae for strategies analyzed and its corresponding error. This value is defined to be equal to the difference between the strategy number of messages and Chord number of messages. Moreover, Figure 2 shows how the number of messages is reduced by SecureRoutingDHT due to usage of the reputation mechanism. Note that results are roughly equivalent to Chord, when number of peers grows.

2) Maintenance: The expected number of required messages during the maintenance process is derived from sum up each message involved in the update of the routing table. On the one hand, the number of expected messages to obtain the corresponding *quorum* is $(h-1)+(2r+1)$. On the other hand, the number of expected routing contacts depends on the underlying DHT, say, C_f, hence the expected amount of messages is:

$$C_f \cdot (h+2r) \tag{4}$$

When Chord is the underlying DHT, roughly $C_f = h = \log_2 n$. Moreover, if $2r = \log_2 n$, as in Myrmic [34], the expected number of messages is $2(\log_2 n)^2$, namely, the complexity is $O((\log_2 n)^2)$. It is important to notice that $2r = \log 2n$ is a value that increments failure tolerance, so it is an acceptable value.

3) Tolerance to malicious peers: This subsection compares the proposed protocol with other approaches regarding the probability of success when a message is forwarded, namely, the measure of its reliability.

From threat model, it can be seen that a peer is malicious with a probability at most f. Hence that probability of E_1, the event representing a path with h hops and not containing malicious nodes, is

$$Pr(E_1) = (1-f)^h \tag{5}$$

Let us consider a multi-path based strategy in where a message is sent through d independent paths. Furthermore, let X_2 be a random variable that represents the number of paths that does not contain any malicious nodes. Therefore, the failure probability of a multi-path based strategy is given by *Pr(fail)≤Pr(X_2=0).*

As it is known that a free-malicious path has probability $Pr(E_1) = (1-f)^h$, then the probability that a path contains at least a malicious node is $1 - Pr(E_1)$. Therefore, the probability of each path would be non-free-malicious is given by $(1 - Pr(E_1))^d$. Finally, the probability that at least a path is free-malicious, $Pr(E_2)$, is given by :

$$Pr(E_2) = Pr(0 \leq X_2) = 1 - (1-(1-f)^h)^d \tag{6}$$

Conversely in wide-path-based strategies, a message is successfully forwarded if at least one peer within each intermediate *quorum* is not malicious. Let E_3 be the event that one message has been forwarded successfully. It is clear that the probability that, in a *quorum* of size d, will be there at least one non-malicious peer is $1 - f^d$. Therefore,

$$Pr(E_3) = (1-f^d)^h \tag{7}$$

For our protocol, analysis is based on that introduced in [26]. Let α be the probability that the reputation mechanism excludes an honest peer and β be the probability that the reputation mechanism chooses a malicious peer. Furthermore, let D_i be a set of size d and E_4 be the event that a malicious peers is selected from a set D_i of size d. Finally, let E_5 be the event that a peer is selected from a quorum by the reputation systems.

Evidently $Pr(E_4) = \frac{f \cdot d}{d} = f$ and $Pr(E_5) = (1-\alpha)(1-f) + f\beta$, where $(1-\alpha)(1-f)$ and $f\beta$ are the probabilities of choosing a honest and malicious peer by the reputation system respectively. Consequently, the probability of the event of choosing a malicious peer in the set D_i given that the reputation system has already chosen one, namely, $\gamma = Pr(E_4|E_5)$ is:

$$\gamma = \frac{Pr(E_4 \cap E_5)}{Pr(E_5)} = \frac{f\beta}{(1-\alpha)(1-f)+f\beta} \tag{8}$$

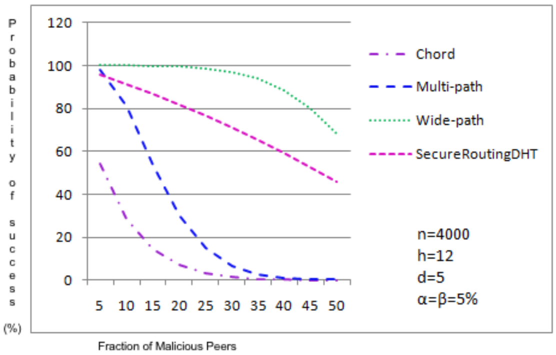


Figure 3. Probabilities of success

As it can be noted, $1 - \gamma$ represents the probability of choosing a non-malicious peer within D_i. Thus, the probability of success of SecureRoutingDHT is $(1 - \gamma)^{h-1}$. This new equation is similar to equation (5) and as can be seen, γ must be smaller than f in order to increment the probability of success. Suppose that $\gamma < f$ and $0 < f, \alpha, \beta < 1$, then :

$$\begin{aligned} \frac{f\beta}{(1-\alpha)(1-f)+f\beta} &< f \\ \frac{\beta}{(1-\alpha)(1-f)+f\beta} &< 1 \\ \beta - f\beta &< (1-\alpha)(1-f) \\ \beta(1-f) &< (1-\alpha)(1-f) \\ \beta &< 1-\alpha \end{aligned} \tag{9}$$

The above means that whether β is sufficiently small, the probability of success of SecureRoutingDHT will increment. For instance, setting to $f = 0,25$, $n = 4000$, $h = \log_2 n$ and $\alpha = \beta = 0,05$, the probability of success of SecureRoutingDHT is 82%. Figure 3 shows the probability of success of the strategies analysed.

As it already has been shown, strategies based on wide-path have a higher probability of success. However, the number of sent messages is higher than other strategies (Figure 2). In turn, the introduced protocol provides an acceptable probability of success while using a smaller number of messages, moreover it can tolerate theoretically up to a fraction of 35% of malicious peers.

B. Simulation

Simulations were performed by using Overlay-Weaver [31]. These were carried out during a period of time, where relevant information was collected in order to measure the number of messages and the probability of success.

The deployed scenarios to evaluate the protocol are described below.

1) *Scenario 1 (scalability)*: Deploying up to 4000 nodes and issuing of requests for a selected key in order to compute the average of number of messages per query.
2) *Scenario 2 (tolerance)*: Deploying 4000 nodes and uniformly distributing a fraction f of malicious peers, in order to evaluate the probability of success if f is incremented.

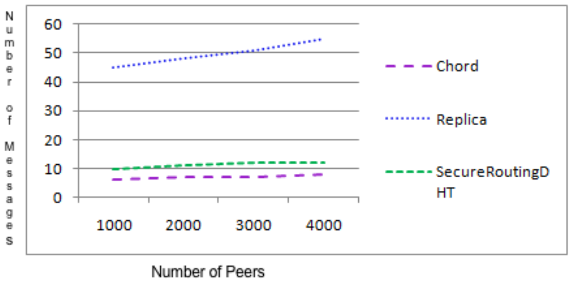


Figure 4. Number of messages in the simulation

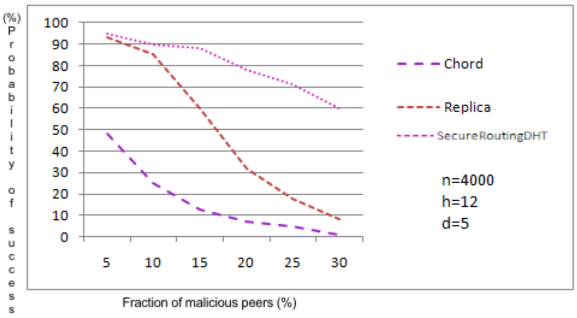


Figure 5. Probability of success in the simulation

Next results are presented according to the scalability and tolerance of malicious peers.

Scalability: The test was performed by choosing random peers and a key k over the system. Each random chosen peer issues a request for the key k and finally the average of messages per query is computed. Figure 4 shows obtained results for Chord, Replica and SecureRoutingDHT. As can be noted, results support the scalability of SecureRoutingDHT in terms of the number of messages.

Tolerance to malicious peers: The test was performed by uniformly distributing a fraction f of malicious peers over the system, namely, $f \cdot 4000$ peers are randomly chosen and considered as malicious. In this scenario a malicious peer does not cooperate with the routing process.

Figure 5 shows results that are obtained for Chord, Replica and SecureRoutingDHT. As it can be noted, the introduced protocol exceeds the probability of success than those guaranteed by Chord and Replica and to tolerate up to 30% of malicious peers, which is an acceptable value due to the fact that solutions to the*Sybil* attack try to limit the number of misbehaivours peers.

VI. Conclusions and Future Work

P2P systems were created without any security considerations; thus, there are a lot of attacks against them, such as *sybil*, eclipse, routing, storage and retrieval attacks. As the routing process is one of the most important mechanisms within the context of P2P systems, this paper addressed threats against this process.

The paper introduces SecureRoutingDHT, a reliable routing protocol that aims to mitigate routing attacks and provide direct access to all replicas of a requested resource. This protocol is compatible with several solutions to the *sybil* attack and it is decoupled from the underlying P2P DHT-based system. As well as reduces the number of messages in comparison with those consumed in S-Chord [10], Replica [11] and Quorum [36].

Furthermore, a theoretical and practical (through simulations) analysis of the protocol are presented, concerning its scalability in terms of number of sent messages and tolerance to malicious peers. Particularly, when SecureRoutingDHT is built on top of Chord, it was theoretically shown that the expected number of messages is $\log_2 n - 1 + |D|$, as well as that the expected number of sent messages during the maintenance protocol is $\log_2 n \cdot (\log_2 n + 2r)$. The above evidences the dependency to churn rates. Finally, as for the reliability, the benefits that were obtained are significant, since that our protocol behaves fairly well up to a 30% percentage of malicious nodes.

Finally, it would be interesting to evaluate performance and probability of success of proposed protocol whenever *iterative* routing is implemented. Additionally, consider other possible mechanisms to obtain recommendations, indeed, there can be taken advantage of back-pointer information for enriching the recommendation process. There is a need for an implementation of this protocol, as well as a set of software libraries, in order that there could be built new applications that take advantage of it.

References

[1] I. Baumgart and S. Mies, ”S/Kademlia: A practicable approach towards secure key-based routing”, Proc. of the 13th Int. Conf. on Parallel and Distributed Systems, IEEE Press, 2007, pp. 1-8, doi: 10.1109/ICPADS.2007.4447808.

[2] K. Butler, S. Ryu, P. Traynor, and P. McDaniel, ”Leveraging identity-based cryptography for node ID assignment in structured P2P systems”, Transactions on Parallel and Distributed Systems, IEEE Press, 2009, pp. 1803-1815, doi: 10.1109/TPDS.2008.249.

[3] H. Cai, J. Wang , D. Li, and J. Deogun ”A novel state cache scheme in structured P2P systems”, Journal of Parallel and Distributed Computing, Academic Press, 2005, pp. 154-168, doi: 10.1016/j.jpdc.2004.09.005.

[4] M. Castro, P. Druschel, A. Ganesh, A. Rowstron, and D. Wallach, ”Secure routing for structured peer-to-peer overlay networks”, Proc. of the 5th Symposium on Operating Systems Design and Implementation, ACM Press, 2002, pp. 299-314, doi:10.1145/844128.844156.

[5] C., Chan, S. Chan, ”Distributed Hash Tables: Design and Applications”, Handbook of Peer-to-Peer Networking, Springer Science, 2010, p. 257-280, doi:10.1007/978-0-387-09751-0_10.

[6] T. Condie, V. Kacholia, S. Sankararaman, J. Hellerstein and P. Maniatis, ”Induced Churn as Shelter from Routing-Table Poisoning”, Proc. of the 13th Symposium on Network and Distributed System Security, 2006.

[7] G. Danezis, C. Lesniewski-Laas, M. Kaashoek, and R. Anderson, ”Sybil-resistant DHT routing”, Proc. of the 10th European Symposium On Research In Computer Security, Springer, 2005, pp. 305-318, doi:10.1007/11555827_18.

[8] J. Douceur, ”The sybil attack”, Revised Papers from the 1st International Workshop on Peer-to-Peer Systems, Springer, 2002, pp. 251-260, doi: 10.1007/3-540-45748-8_24.

[9] N. Fedotova, M. Bertucci, and Veltri, ”Reputation management techniques in DHT-based peer-to-peer networks”, Proc. of the 2nd Int. Conf. on Internet and Web Applications and Services, IEEE Press, 2007, pp. 4, doi: 10.1109/ICIW.2007.53.

[10] A. Fiat, J. Saia, and M. Young, ”Making chord robust to byzantine attacks”, Proc. of the 13th Annual European Symposium on Algorithms, Springer, 2005, pp. 803-814, doi: 10.1007/11561071_71.

[11] C. Harvesf and D. Blough, ”Replica placement for route diversity in tree-based routing distributed hash tables”,Transactions on Dependable and Secure Computing, IEEE Press, 2009, doi: 10.1109/TDSC.2009.49.

[12] K. Hoffman, D. Zage, and N. Nita-Rotaru, ”A survey of attack and defense techniques for reputation systems”, Computing Surveys, ACM, 2009, pp. 1-31, doi: 10.1145/1592451.1592452.

[13] A. Jøsang, R. Ismail, and C. Boyd, ”A survey of trust and reputation systems for online service provision”, Decision Support Systems, Elsevier Science Publishers, 2007, pp. 618-644, doi: 10.1016/j.dss.2005.05.019.

[14] A. Jøsang, and W. Quattrociocchi, ”Advanced features in bayesian reputation systems”, Proc. of the 6th Int. Conf. on Trust, Privacy & Security in Digital Business, Springer, 2009, pp. 105-114, doi: 10.1007/978-3-642-03748-1_11.

[15] A. Kapadia and N. Triandopoulos, ”Halo: High-assurance locate for distributed hash tables”, Proc. of the 15th Symposium on Network and Distributed System Security, 2008.

[16] F. Lesueur, L. Mé, and V. Triem Tong, ”A sybilproof distributed identity management for P2P networks”, Proc. of Symposium on Computers and Communications, IEEE, 2008, pp. 246-253, doi: 10.1109/ISCC.2008.4625694.

[17] B. Levine, C. Shields, and B. Margolin, ”A Survey of Solutions to the Sybil Attack”, Technical Report, University of Massachusetts, 2006.

[18] S. Marti, P. Ganesan and H. Garcia-Molina, ”DHT routing using social links”, Revised Selected Papers from the 3rd International Workshop on Peer-to-Peer Systems, Springer, 2004, pp. 100-111, doi: 10.1007/978-3-540-30183-7_10.

[19] M. Naor and U. Wieder ”A simple Fault Tolerant Distributed Hash Table”, Revised Papers from the 2nd Int. Workshop on Peer-to-Peer Systems, Springer, 2003, pp. 88-97, doi: 10.1007/978-3-540-45172-3_8.

[20] K. Needels and M. Kwon, ”Secure routing in peer-to-peer distributed hash tables”, Proc. of the Symposium on Applied Computing, ACM, 2009, pp. 54-58, doi: 10.1145/1529282.1529292.

[21] B. Roh, O. Kwon, S. Hong, and J. Kim, ”The exclusion of malicious routing peers in structured P2P systems”, Proc. of the 5th Int. Workshop on Agents and Peer-to-Peer Computing, Springer, 2006, pp. 43-50, doi: 10.1007/978-3-540-79705-0_4.

[22] C. Roncancio, M. Villamil, C. Labbé, and P. Serrano-Alvarado, ”Data Sharing in DHT Based P2P Systems”, Transactions on Large-Scale Data- and Knowledge-Centered Systems, Springer, 2009, pp. 327 - 352, doi: 10.1007/978-3-642-03722-1_13.

[23] M. Sánchez-Artigas, P. García-López, and A. Gómez, ”A novel methodology for constructing secure multi-path overlay”, Internet Computing, IEEE Press, 2005, pp. 50-57, doi: 10.1109/MIC.2005.117.

[24] M. Sánchez-Artigas, P. García-López, and A. Gómez, ”By-pass: providing secure DHT routing through bypassing malicious peers”, Proc. of Symposium on Computers and Communications, IEEE Press, 2008, pp. 934-941, doi: 10.1109/ISCC.2008.4625618.

[25] M. Sánchez-Artigas and P. García-López, ”On routing in distributed hash tables: is reputation a shelter from malicious behavior and churn?”, Proc. of the 9th Int. Conf. on Peer-to-Peer Computing, IEEE Press, 2009, pp. 31-40, doi: 10.1109/P2P.2009.5284546.

[26] M. Sánchez-Artigas, P. García-López, and A. Gómez, ”Secure forwarding in DHTs-is redundancy the key to robustness?”, Proc. of the 14th International European Conference on Parallel and Distributed Computing, Springer, 2008, pp. 611-621, doi: 10.1007/978-3-540-85451-7_65.

[27] A. Singh, T. Ngan, P. Druschel, and D. Wallach, ”Eclipse attacks on overlay networks: Threats and defenses”, Proc. of the 25th Int. Conf. on Computer Communications, IEEE Press, 2006, pp. 1-12, doi: 10.1109/INFOCOM.2006.231.

[28] E. Sit and R. Morris ”Security considerations for peer-to-peer distributed hash tables”, Revised Papers from the 1st Int. Workshop on Peer-to-Peer Systems, Springer, 2002, pp. 261-269, doi: 10.1007/3-540-45748-8_25.

[29] M. Srivatsa and L. Liu, ”Vulnerabilities and security threats in structured overlay networks: A quantitative analysis”, Proc. of the 20th Annual Computer Security Applications Conference, IEEE Press, 2004, pp. 252-261, doi: 10.1109/CSAC.2004.50.

[30] I. Stoica, R. Morris, D. Karger, M. Kaashoek, and H. Balakrishnan, ”Chord: A Scalable Peer-to-peer Lookup Service for Internet Applications”, Proc. of the Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications, ACM, 2001, pp. 149-160, doi: 10.1145/383059.383071.

[31] K. Shudo, Y. Tanaka, and S. Sekiguchia, ”Overlay Weaver: An overlay construction toolkit”, Computer Communications, Elsevier, 2008, pp. 402-412, doi: 10.1016/j.comcom.2007.08.002.

[32] G. Urdaneta, G. Pierre, M. Van Steen, ”A survey of DHT security techniques”, Journal ACM Computing Surveys, ACM Press, 2011, Volume 43 Issue 2, doi:10.1145/1883612.1883615.

[33] R. Villanueva, M. Villamil and R. Arnedo, ”Secure routing strategies in DHT-based systems”, Proc. of the Third Int. conf. on Data management in grid and peer-to-peer systems, Springer, 2010, pp. 62-74, doi:10.1007/978-3-642-15108-86.

[34] P. Wang, L. Osipkov, N. Hopper, and Y. Kim, ”Myrmic: secure and robust DHT routing”, Technical report, University of Minnesota-Twin cities, 2006.

[35] X. Xiang, and T. Jin, ”Efficient secure message routing for structured peer-to-peer systems”, Proc. of the Int. Conf. on Networks Security, Wireless Communications and Trusted Computing, IEEE Press, 2009, pp. 354-357, doi: 10.1109/NSWCTC.2009.124.

[36] M. Young, A. Kate, I. Goldberg, and M. Karsten, ”Practical robust communication in DHTs tolerating a byzantine adversary”, Proc. of the 30th Int. Conf. on Distributed Computing Systems, IEEE Press, 2010, pp. 263 - 272, doi:10.1109/ICDCS.2010.31.

Domain-based Testbed for Peer-to-Peer Information Retrieval

Saloua Zammali
Dept. of Computer Science
and Mathematics
Faculty of Sciences
of Tunis, Tunisia
Email: zammalisalwa@gmail.com

Amira Ben Salem
Dept. of Computer Science
and Mathematics
Faculty of Sciences
of Tunis, Tunisia
Email: bnsalemamira@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*—Information retrieval (IR) is a field that deals with storage and access to relevant information according to the user needs. The main goal of an Information Retrieval System (IRS) is to return to the user the most valuable documents in response to his queries. Classical models in IR are based on a general approach that meets the users invariably returning the same results for two users with the same issued query but having different information needs and different research preferences. Hence, the need to combine user domain interests with information retrieval becomes a challenge. The major issues raised by information retrieval, mainly, concerns domain interests modeling and domain exploitation in IR models. The major limitation of testbeds in distributed information retrieval (DIR) is mainly related to testbed that does not include domain interests as a source of evidence for evaluation of relevant documents. This problem becomes more insistent in Peer-to-Peer Information Retrieval (P2PIR) where there is not yet a standard testbeds for use. In this paper, we propose, *DBT*, a Domain-based Testbed for P2PIR. DBT is based on a new method for modeling peer and query domains. We represent these domains by using YAGO ontology.**

***Keywords-Testbed*; *Information retrieval*; *P2P systems*; *YAGO ontology*.**

I. Introduction

Information retrieval (IR) is a field that deals with storage and access to relevant information according to the user needs. The main goal of an information retrieval system (IRS) is to return to the user the most valuable documents in response to his queries. For a user query, an IR system allows to find a subset of potentially relevant documents, from a documents collection, responding to this query.

The growth of the web has delivered the IR face new challenges of access to information, namely to find relevant information in a diversified area and considerable size and that meets the need for specific user information. The major limitation of most classical information retrieval system is that they return, for a same query submitted by different users, the same results. However, users have different search background like interests, preferences, etc.

Studies, in [1], show that the problem of these systems lies partly in the fact that they are based on a general approach that considers the user information needs is completely represented by its query. To overcome this issue, the representation of user need must be extended in order to return the most useful information. As a result, the evaluation methodologies of these systems have been challenged by the consideration of extra external knowledge rather than the queries terms. That's why an appropriate testbed is needed, either in centralized IR or in distributed IR (particulary P2PIR) where queries and peers have limited descriptions. The testbed will be extended by semantic information provided from a semantic resource such as ontologies.

The main purpose of this extension is to make the testbed more enriched where user (*i.e.*, peer) need is not only represented by his queries but also through domains that describe the subject of the queries and peers. In this paper, we propose a domain-based testbed, suitable for the evaluation of P2PIR systems that takes in consideration the domain of queries and peers.

The paper is organized as follows. In Section II, we recall the key notions used throughout this paper. We review, in Section III, related work about building testbed for IR. In Section IV, we describe our approach of building distributed domain-based testbed. In Section V, we show our first experimental results. Finally, we present our conclusions regarding the current work and how this may relate to future trends P2PIR systems.

II. Key notions

Before presenting our approach, we provide a simplified definition for some of the key concepts used throughout in this paper, namely, *testbed* and *ontology*.

A. Notion of ontology

An ontology represents knowledge as a set of concepts within a domain, and the relationships between those concepts. It can be used to reason about the entities within that domain and may be used to describe the domain [2]. An ontology can be constructed in two ways, domain dependent and generic. CYC [3], WordNet [4], and Sensus [5] are examples of generic ontologies.

One way of introducing external knowledge into IR is by using ontology, for instance, by means of a list of keywords that reflect knowledge about the domain.

B. Notion of testbed

Definition 1 (Centralized information retrieval testbed): **Testbed** = documents collection + queries collection + relevance judgments.
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 [6].

Definition 2 (Distributed information retrieval testbed): **Testbed** = documents collection+ queries collection + documents and queries distribution method among peers + documents and queries replication method among peers + evaluation metrics+ queries responses [7].

Definition 3 (Domain-based DIR testbed): We define the main components which a domain-based distributed testbed must provide as follows:

1) Test collection: documents collection, queries collection and relevance judgments.
2) A definition of a documents and queries distribution methods among peers.
3) A definition of a documents and queries replication methods among peers.
4) A set of semantic resources, such as ontologies, which provide semantics information.

In the following section, we review various work on the building testbeds in a centralized and distributed systems.

III. Related work on testbed building for DIR

A. Testbeds for centralized systems

For centralized Information Retrieval, there exist a significant number of standard centralized testbeds, such as the yearly competitions conducted by Cranfield [8] TREC [9], DMOZ [10], etc.

- Cranfield: Cranfield is the first centralized testbed, was created under the direction of C. Cleverdon. It is started in 1957 [8]. Cranfield is composed of 1400 documents and 221 queries [8].
- TREC: Text REtrieval Conference(TREC) [9] is designed as a series of workshops in the field of information retrieval.

B. Testbeds for decentralized systems

Building testbeds for distributed information retrieval systems is a challenge, in particular in P2PIR systems. Indeed, there is not yet a standard testbeds for use. To overcome this lack, Peer-to-Peer Information Retrieval benchmarking (P2PIRB), a framework for building distributed testbeds, is proposed in our previous studies [6]. P2PIRB framework provides a certain nombre of testbeds (such as Uniform Testbed, Random Testbed and specialized Testbed) [6].

C. Summary

Testbeds for classical centralized/decentralized IR systems have several problems, among which we can mention:

1) Testbeds are based on queries which are the only resources reflect information needs key of the user. Indeed, information needs of the user is represented by a single key resource, including a query keywords often expressed in natural language.
2) The interests of users, having made these queries, does not form a part of testbed.
3) Absence of real users: traditional evaluation model does not include real users in research contexts and replaces them with experts responsible for creating relevance judgments for each topic.
4) Classical evaluation measures are not exhaustive in the sense that the document is considered relevant if it recovers query topic, independently of the context and the task of research.

In this paper, we focus on the two first limits. In the literature, few approaches have been proposed for integration of interest domains in centralized testbeds [11][12][13]. However, theses testbeds are not freely available and not standardized. To the best of our knowledge, building domain-based testbeds has not been widely addressed in distributed information retrieval.

To tackle this limitations of traditional testbeds, in recent years, there has been an increasing research interest in the problem of enrichment testbed with domains of interest. Addressing these issues, we propose a domain-based distributed testbed suitable for the evaluation of P2PIR systems.

IV. Domain-Based Testbed for P2PIR

A. Global architecture of creating a domain-based distributed testbed

The aim of our approach is to build a distributed testbed extended with metadata representing the domains of query and peer. The use of domain in evaluation approaches addresses the above limitations of the traditionnal evaluation. Therefore, the proposed approach consists of three parts: testbed building, domain modeling and domain integration. The architecture of the process of creating a domain-based distributed testbed is described in Figure 1.

B. Testbed building

To distribute documents and queries among the set of peers, we used the Benchmarking Framework for P2PIR [6]. This framework is configurable, which allows user to choose certain parameters (*i.e.*, number of peers, replication of queries, etc.) and provides XML files describing the nodes, the associated documents and the queries to be launched on the network.

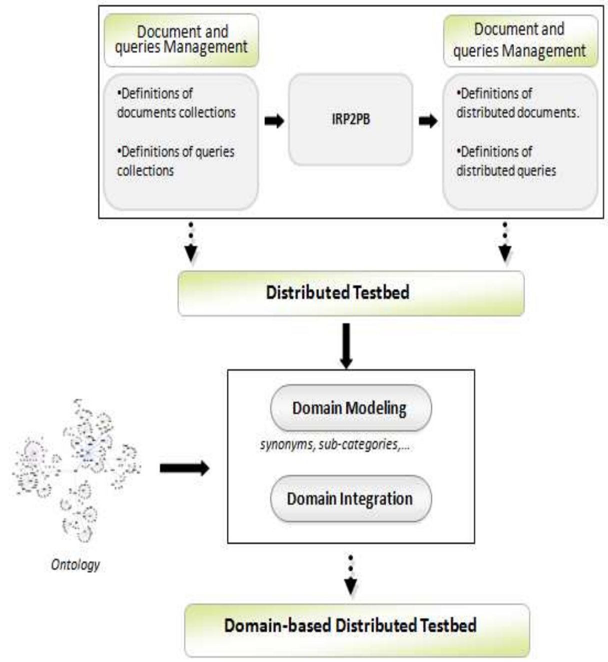

Document and queries Management
•Definitions of documents collections
•Definitions of queries collections
IRP2PB
Document and queries Management
•Definitions of distributed documents.
•Definitions of distributed queries
Distributed Testbed
Domain Modeling
synonyms, sub-categories,...
Domain Integration
Ontology
Domain-based Distributed Testbed

Figure 1. Global architecture of creating a domain-based distributed testbed

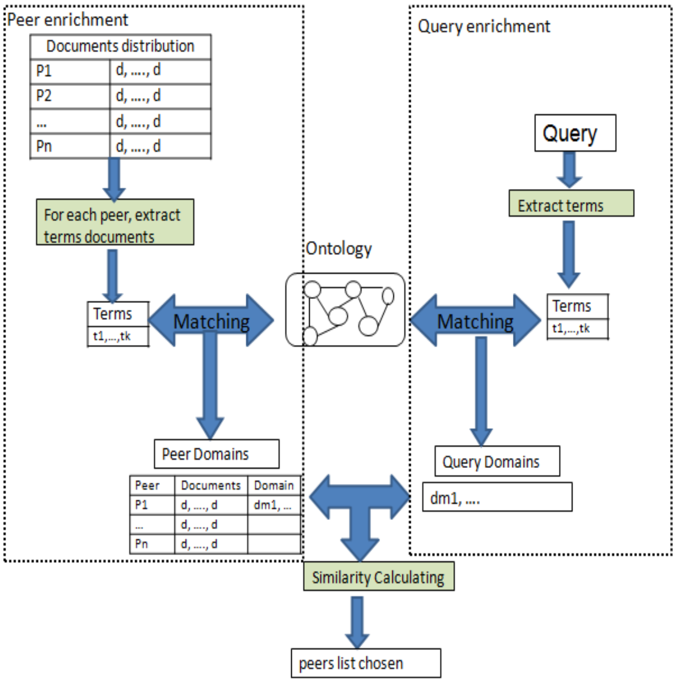


Figure 2. Domain modeling process

C. Ontology-based domain modeling

Modeling domain, in our work, is based on extracting knowledge from a given ontology. Knowledge extraction refers to the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources such as text. Our domain modeling can be formulated as follows:
Let $\mathcal{T} = \{t_1, \dots, t_n\}$: a set of terms and O: an ontology.
After doing the correspondence between the terms of $\mathcal{T}$ and the entities of $\mathcal{O}$, we obtain, from $\mathcal{O}$, a set of entities (*i.e.*, terms). For several terms, we will obtain a larger set of ontology entities. The entities with higher frequency are selected and it represents the most appropriate sense to the set of terms: $\mathcal{E} = \{e_1, \dots, e_i\}$.
For each e_i, we extract from $\mathcal{O}$:

- a set of synonyms $Syn = \{s_1, \dots, s_j\}$,
- a set of general terms $\mathcal{G} = \{t_{g1}, \dots, t_{gk}\}$,
- a set of specific terms $\mathcal{S} = \{t_{s1}, \dots, t_{sl}\}$.

Therefore, the domain $\mathcal{D}$, of the set of terms in $\mathcal{T}$, is represented by a set of entities $\mathcal{E}$, called domains. Each domain e_i, is represented by the set of synonyms *Syn*, the set of general terms $\mathcal{G}$ and the set of specific terms $\mathcal{S}$: $\mathcal{D} =< e_i, Syn, \mathcal{G}, \mathcal{S} >$

D. Domain integration in testbed

A node, in a P2P network, contains a collection of homogeneous documents that represents its center of interest. In order to realize this, we use a dataset that reflects real scenarios. Figure 3 illustrates the structure of a peer content. Each peer is described by its documents and a set of domains:

$$p = (docs, \{dom_1, dom_2, \dots, dom_k\})$$

where: $docs$ is the documents of peer p and k is the number of domains for considered peer and each domain is constructed as follows:

$$dom = (synos, sub\_catgs, sous\_catgs)$$

$synos$, sub_catgs, $sous\_catgs$ are respectively: synonyms, sub categories, sous categories of peer documents.
Figure 2 provides a visual representation of domain-based peer enrichment.

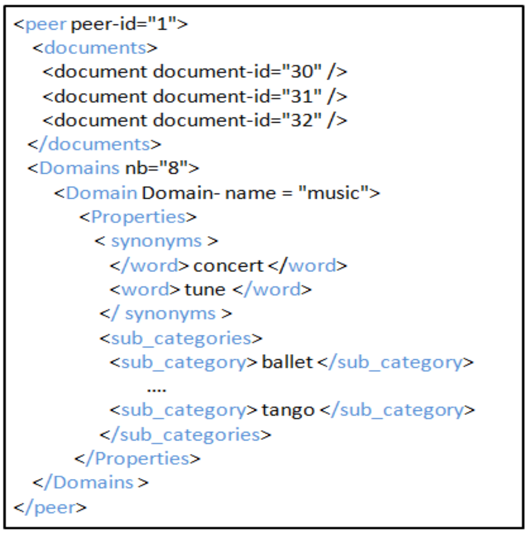
V. Evaluation Methodology

In order to get meaningful terms in the centralized collection, we decided to build a new test collection, where queries and documents terms are not generated randomly but in a way to ensure semantic between terms.
To build a test collection, you must specify:

- What are the criteria for the selection of documents.
- How to identify relevant documents for each query.

A. Document collection

We choose to use Delicious [14] tags and we consider the tags made by each user for a specified article as the terms of a document. For this purpose, we used the dataset published in Social-ODP-2k9 Dataset [15].



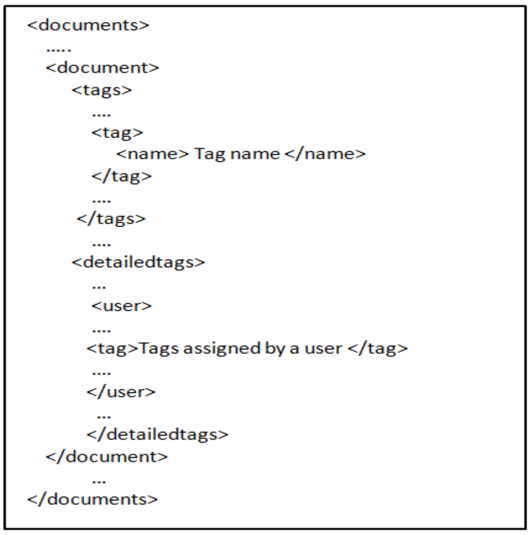
```
<peer peer-id="1">
  <documents>
    <document document-id="30" />
    <document document-id="31" />
    <document document-id="32" />
  </documents>
  <Domains nb="8">
     <Domain Domain- name = "music">
        <Properties>
          < synonyms >
            </word> concert </word>
            <word> tune </word>
          </ synonyms >
          <sub_categories>
            <sub_category> ballet </sub_category>
                ....
            <sub_category> tango </sub_category>
          </sub_categories>
        </Properties>
  </Domains >
</peer>
```

Figure 3. Illustration of the peer's structure

Delicious is a website that save and share tagged web page and classify them according to the principle of folksonomy by tags. It was created in 2003 by Joshua Schachter in order to save their personal bookmarks. The site interface is based on HTML, which makes the site easy to use. Delicious content is organized via RSS (Rich Site Summary) and is based on *tags* notion. Tags are keywords describing the content of the document (*e.g.*, Sports, Cinema, Internet, etc.). We investigate the tag sets in delicious thanks to both its popularity and availability. Social-ODP-2k9 is a dataset created during December 2008 and January 2009 with data retrieved from Delicious and StumbleUpon social bookmarking sites, the Open Directory Project and the Web. It is available for research purposes and has XML format, as shown in Figure 4. The tags $<document>$ and $</document>$ mark the beginning and the end of a document respectively, and each document has a number of users (encapsulated in a $<user>$ element) who have tagged (delimited by $<tag>$ element). The construction of test collection from delicious is described by algorithm 1. All articles (*i.e.*, documents) in AC collection is partitioned according to the number of documents per peers. To build the documents collection associated to peer i, we use $ExtractTagsFromUsers$ algorithm to extract the tags corresponding to the article i.

Algorithm 1: DOCUMENTS-BUILD

```
Algorithm: Documents-Build($AC$, $Pnb$, $n$)
Input:
$AC$: Articles collection.
$Pnb$: Peers number in network.
$n$: Documents number per peers.
Output:
$\mathcal{DF} = \{D_{peer_1}\} \cup \{\ldots\} \cup \{D_{peer_{Pnb}}\}$: Documents collection.
begin
    for ($i = 1$; $i < |AC|$; $i{+}{+}$) do
        $D_{peer_i} := \emptyset$;
        for ($j = 1$; $j < n$; $j{+}{+}$) do
            $d_j$ = ExtractTagsFromUsers($j$, $D_{Pnb}$);
            $D_{peer_i} = D_{peer_i} \cup \{d_j\}$;
    $\mathcal{DF} = \{D_{peer_1}\} \cup \{\ldots\} \cup \{D_{peer_{Pnb}}\}$ ;
    return ($\mathcal{DF}$)
```



```
<documents>
  .....
  <document>
     <tags>
       ....
       <tag>
          <name> Tag name </name>
       </tag>
       ....
     </tags>
       ....
     <detailedtags>
       ...
       <user>
       ....
       <tag>Tags assigned by a user </tag>
       ....
       </user>
       ...
       </detailedtags>
  </document>
       ...
</documents>
```

Figure 4. Illustration of the delious's structure

Delicious is based on tags technology. Tags are in the form of a word (*e.g.*, sports, movies, Internet, etc.) can quickly find relevant sites to the tag. Therefore, an ODP site (having url) can be tagged by multiple users with different terms. In our case, we considered:

- The URLs of ODP represent peers.
- The tags, for a given user and article, represents document terms.

The idea behind this choice is that each peer usually has a homogeneous collection of documents representing these interests. However, an ODP article is tagged by several users, but these tags, necessarily, have a certain correlation between them. To simulate this behavior and remain in a realistic environment, we have assigned the sets of tags (each set of tags represents a document), corresponding to a given URL, to a given peer.

B. Queries collection and relevance judgements

A query represents the user information need. Queries collection must adequately model human users behavior. Indeed, queries collection should represent the needs of non-expert users (for example, ambiguous query represented by a single term) and must also represent expert need users.

Studies have shown that the queries submitted by users are relatively short and are generally limited to less than three keywords [16]. For this, we have established three set of queries: the first contain one term, the second contain two terms and the third contain three terms.
Relevance judgements are obtained using the cosine function, given in equation 1.

$$S(d_j, q) = \cos(\overrightarrow{q}, \overrightarrow{d_j}) = \frac{d_j \times q}{|d_j| \times |q|} \qquad (1)$$

The cosine function, given in equation 1, is often used to determine the similarty between a document d_j and a query q.

C. Queries Distribution

Queries distribution among peers is done in a completely random manner, but under the constraint that queries repartition is proportional to the documents one. We used the *IRP2PB* tool for queries distribution on peers [6].

D. Ontology

YAGO (Yet Another Great Ontology) is a huge semantic knowledge base. Figure 5 represent a fragment of YAGO knowledge representation. It contains 2 million entities (such as persons, organizations, cities, etc.). This ontology contains 20 million facts about these entities [17]. The main reasons for using this resource are:

- It is derived from Wikipedia and WordNet.
- It exists in many formats (XML, SQL, RDF, etc.).
- It covers a vast amount of individuals.

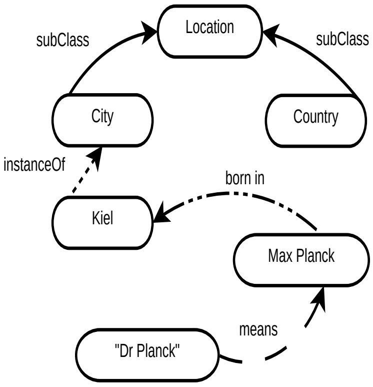


Figure 5. Fragment of YAGO ontology

E. PeerSim simulator:

To evaluate the approach proposed in this paper, we have chosen to use the PeerSim [18] simulator which is an open source tool written in Java. It has the advantage of being dedicated to the study of P2P systems [7]. 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 [19]. This extension can be seen as a PeerSim specialization for information retrieval.

F. Routing Algorithms:

- *Gnutella*: 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 [6].
- *DBR* (Domain-based routing): The pseudo-code for our routing algorithm is given by algorithm 2.
  The *DBR* peer selection algorithm uses the YAGO ontology for select suitable peers. This is due to the enrichment of peer structure by its interests (*i.e.*, extracted domains from YAGO). Indeed, initially, each peer has a set of documents representing their interests (*i.e.*, domains). In order to express, explicitly, we used YAGO ontology (as detailed previously in the section IV-C).
  For a query Q, the algorithm determines from ontology, a set of domains associated to the query (denoted by *QueryDoms*: getQueryDomains).
  Determine the set of domains, for each pair, denoted by $PeersDocsDoms$ (getPeersDocsDomains() function).
  For each peer domain, determine a set of domains similar to Q which are sorted according to the similarity value (getSimilarDomain() function of algorithm 2).
  The similarity between a domain $dom \in PeersDocsDoms$ and the domains *QueryDoms* of Q is determined by the formula as follows:

$$Sim(QueryDoms, dom) = \frac{|QueryDoms \ \cap \ dom|}{|QueryDoms \ \cup \ dom|} \qquad (2)$$

G. Evaluation Metrics:

To compare the performance of the two routing algorithms, we used the metrics Recall (R) and Precision (P) defined as follow: given a query Q, consider RDR the number of relevant documents returned, RD is the number of relevant documents and DR the number of documents returned:

$$R(Q) = \frac{RDR}{RD} \qquad (3)$$

$$P(Q) = \frac{RDR}{DR} \qquad (4)$$

H. Initialize simulation parameters

The simulation, of both algorithms DBR and Gnutella, is based on the parameters:

- *TTL* (Time To Live): Maximum depth of research, initialized to 4.

- *Pmax*: Maximum number of peers which the query should be propagated to.
- *Overlay size*: Number of peers in the network, initialized to 500.

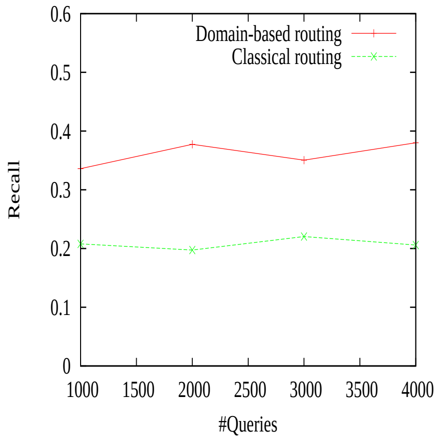


Figure 6. Relation between Recall and Nbr of Queries according to Gnutella and DBR algorithms

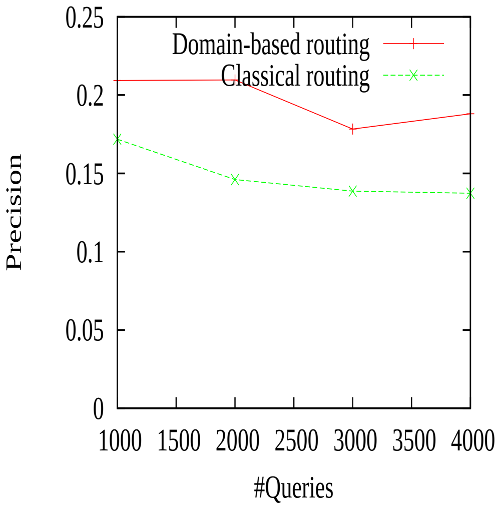


Figure 7. Relation between Precision and Nbr of Queries according to Gnutella and DBR algorithms

I. Experimental Results

In this experiment, we compared the performance of routing algorithms performed with and without considered interest domains of user.

To compare the performance of the domain-based algorithm (*i.e.*, DBR) and the classical routing (*i.e.*, Gnutulla), we calculate the average recall and precision per interval of 2000 queries sent by different peers in the system.

Figure 6 shows that the average recall of DBR algorithm is between 0.33 and 0.38 while the average recall for Gnutella is between 0.19 and 0.22.
Figure 7 shows that the average precision of DBR algorithm is between 0.18 and 0.20 while the average precision for Gnutella is between 0.13 and 0.17.

These results show that the *DBT* testbed significantly improves the effectiveness of the DBR routing algorithm. In addition, when comparing classical routing to the domain-based routing, we see better recall and precision in the search results since domain-based query retrieve documents that would not be retrieved by using only the keyword-based query.

VI. Conclusion and Future Work

The field of information retrieval is very experimental in nature. We identify the need to create testbeds for information retrieval experimentation. We propose, *DBT*, a testbed for P2PIR, based on interests domains peers. In this paper, we demonstrated that ontology can be used to model peer interest domains and these domains can be used to improve distributed information retrieval.

The first tests presented in this paper are very encouraging. One possible perspective to this work is to vary the number of documents and queries and use other routing algorithms in the aim of making *DBT* testbed more used. We plan to study a new dimensions such as peer location, time and integrate them in distributed testbeds to the aim of improving search effectiveness.

Algorithm 2: Domain-Based Routing Algorithm

```
Algorithm: Domain-Based Routing Algorithm($Q$, $O$)
Input:
$Q$: Query.
$O$: Ontology.
Output:
$selectedPeers$ : selected peers list.
begin
    $QueryDoms$ := getQueryDomains($Q$, $O$) ;
    $PeersDocsDoms$ := getPeersDocsDomains($O$);
    $SimQP$ := $\emptyset$;
    foreach $PDom \in PeersDocsDoms$ do
        $SimQP$ := $SimQP \cup$
        getSimilarDomain($PDom$, $QueryDoms$);
    $selectedPeers$ := getSelectedPeers($SimQP$) ;
    return ($selectedPeers$)
```

References

[1] J. Budzik and K. Hammond, “User interactions with every applications as context for just-in-time information access,” in *Proceedings of the 5^{th} international conference on intelligent user interfaces*, Mars 2000, pp. 44–51.

[2] R. Gruber, “Toward principles for the design of ontologies used for knowledge sharing,” *International Journal of Human-Computer Studies*, vol. 43, pp. 907–928, December 1995.

[3] D. B. Lenat, “Cyc: a large-scale investment in knowledge infrastructure,” *Commun. ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[4] G. A. Miller, “Wordnet: a lexical database for english,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[5] B. Swartout, R. Patil, K. Knight, and T. Ross, “Toward distributed use of large-scale ontologies,” in *Proceedings of the* 10^{th} *workshop on knowledge acquisition for knowledge-based systems*, Canada, 1996.

[6] 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.

[7] ——, “An Evaluation of a Cluster-based Testbed,” in 6^{th} *International Conference on Internet and Web Applications and Services*, The Netherlands Antilles, March 2011, pp. 136–141.

[8] C. Cleverdon, “The cranfield test on index language devices,” in *Association of special libraries (Aslib)*, London, March 1967, pp. 173–194.

[9] “Trec web site,” January 2012, http://trec.nist.gov/.

[10] “DMOZ web site,” January 2012, http://www.dmoz.org/.

[11] L. Tamine-Lechani, M. Boughanem, and M. Daoud, “Evaluation of contextual information retrieval effectiveness: overview of issues and research,” *Knowl. Inf. Syst.*, vol. 24, pp. 1–34, July 2010.

[12] A. Sieg, B. Mobasher, and R. Burke, “Web search personalization with ontological user profiles,” in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, Portugal, 2007, pp. 525–534.

[13] A. Sieg, B. Mobasher, S. Lytinen, and R. Burke, “Using concept hierarchies to enhance user queries in web-based information retrieval,” in *Proceedings of the International Conference on Artificial Intelligence and Applications*, 2004.

[14] “Delicious web site,” January 2012, http://delicious.com/.

[15] “Social ODP web site,” January 2012, http://nlp.uned.es/social-tagging/socialodp2k9/.

[16] A. Spink, H. Ozmutlu, S. Ozmutlu, and B. Jansen, “U.s. versus european web searching trends,” *SIGIR Forum*, vol. 36, pp. 32–38, September 2002.

[17] F. Suchanek, G. Kasneci, and G. Weikum, “Yago: a core of semantic knowledge,” in *Proceedings of the* 16^{th} *international conference on World Wide Web*, Canada, 2007, pp. 697–706.

[18] M. Jelasity, A. Montresor, G. P. Jesi, and S. Voulgaris, “The peersim simulator,” `http://peersim.sf.net/`, January 2010.

[19] “RARE project,” January 2012, http ://www-inf.int-evry.fr/defude/RARE/.

Distinguishing Obligations and Violations in Goal-based Data Usage Policies

Sebastian Speiser
Karlsruhe Service Research Institute (KSRI)
Institute of Applied Informatics and Formal Description Methods (AIFB)
Karlsruhe Institute of Technology (KIT), Germany
Email: speiser@kit.edu

***Abstract*—Laws, regulations and contracts often allow actions, such as usages of data artefacts, under the condition that a set of obligations is fulfilled. Formalising such allowances as policies enables the automated testing whether actions are compliant. Previous approaches to formalise obligations treat them as special objects in the underlying logic. We propose to represent obligations and other compliance conditions in a uniform way, in order to increase understandability by non-expert users. The challenge of such an approach is to differentiate between policy violations and not yet fulfilled obligations. We present a solution based on abductive reasoning, which is described in general terms for policy languages based on first-order logic (FOL). Furthermore, we discuss the use of decidable fragments of FOL as a base for practical policy languages.**

***Keywords*-policy; obligations; usage policies**

I. Introduction

Processes that have to comply with laws, regulations, norms, licenses, and contracts are ubiquitous. Data owners can restrict processes that use their data based for example on privacy law or copyright law. One important aspect of such restrictions are obligations. Obligations are duties that have to be fulfilled, when a right is exercised. Consider for example that it is allowed for a blogger to reuse an image in a non-commercial blog post, however the blogger is obliged to add an attribution of the original creator of the image to his post. Here, the obligations are clearly separated from other restriction, e.g., that the post must be non-commercial. This separation is also found in existing approaches to formalise such restrictions as computer-understandable policies, where obligations are modelled with special operators. Formalisations are useful to automate compliance checks in order to enable systems that adhere to restrictions or assist human users to do so. Special obligation operators that specify, which actions have to be performed to reach compliance, have two drawbacks:

- The policy remains at a lower conceptual level than goal-based policies, which only describe compliant states, and leave the computation of required actions to the policy engine.
- Special operators in a language or its underlying logic mean additional effort for non-expert users to understand them and use them correctly.

In this paper, we propose a novel approach to represent policy conditions and obligations in a uniform way as goal-based policies. Definitions of obligations are given specific for each application in the corresponding domain vocabulary, defined and understood by the system users. Our approach is defined and described in abstract terms using first-order logic and abductive reasoning. We also discuss concrete policy languages that can be used to apply the theoretical results to practical problems.

The rest of the paper is structured as follows: in Section II, we introduce a motivating use cases for formalising policies with obligations. Goal-based data usage policies are introduced in Section III. Our core approach is explained in Section IV. Throughout the technical parts, we go through one continuous example from the use case to illustrate the introduced concepts. In Section V, we discuss practical policy languages and the realisation of the use case. Finally, we discuss related work in Section VI, and conclude in Section VII.

II. Use Case: Restricted Data Usage

Usage of data artefacts can be restricted on the foundation of copyright and privacy laws, company internal guidelines or social norms. We consider usage policies of data artefacts as the formal specification, which usages are allowed and which conditions apply. For formalizing policies, we need a vocabulary for describing data usages, which is visualised in Figure 1. The vocabulary describes Artefacts that can be used in Processes. An artefact has a Policy, to which processes using the artefact must comply. Processes are divided in (i) Usages, which consume an artefact for a specific Purpose, and (ii) Derivations, which generate new artefacts on the base of the used artefacts. A process a can trigger another process b, meaning that the execution of a will lead to the execution of b. We present the following examples of conditions on allowed data usages:

- A derivation of an artefact with usage policy p is allowed, if the generated artefact will also be assigned the same policy p. Such conditions are used, e.g., in Creative Commons ShareAlike licenses.

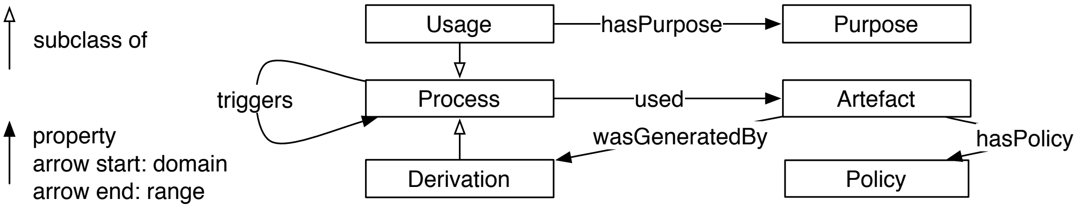


Figure 1. Vocabulary for Data Usage Restrictions

- A usage of an artefact is allowed, if the usage is for non-commercial purposes and triggers an attribution of the artefact's creator. Such conditions are used, e.g., in Creative Commons NonCommercial (NC), Attribution (BY) licenses (abbreviated as BY-NC).
- A usage of an artefact (e.g., a electronic health record) is allowed by doctors, who can also store the artefact under the condition that it is deleted after one year.

III. Goal-based Data Usage Policies

In computer science, the notion of a *policy* refers to a formal description of the actions and behaviors that are allowed or required in a protected context. The context can be characterised for example by the data artefacts that are used, properties or identities of agents involved in performing the action, or temporal constraints. Formal specifications enable the automated detection of policy violations of systems or agents that are formally described. Additionally, in many applications, required adoptions to transit from violation to compliance, can be automatically computed and realised by the corresponding system or agent. In this sense, policies can be used to formalise laws, norms and regulations that apply to a computer system, or a process realised or supported by such a system.

In our approach, we consider goal-based policies as defined by Kephart and Walsh [1]. Goal-based policies are on a high conceptual level, as they only describe the desired states of (the modelled) world, instead of specifying how such a state can be reached.

In the following, we give a general formal definition of policies based on first-order logic (FOL). We assume that the state of the world is described by the FOL theory T of the signature consisting of constants C and predicates P. A policy p is applicable to a set $S_p \subseteq C$ of policy subjects. Policy subjects can either be compliant or non-compliant with p, all other constants $c \in C \setminus S_p$ are called inapplicable. A policy p is defined by a formula $\phi_p[x]$ with one free variable. The compliant subjects are given by the set of constants that when replacing x in ϕ establish $T \models \phi$.

In the following, we restrict all given theories, formulae, and policies to stay in the Datalog fragment of first-order logic. Datalog is the FOL language of function-free Horn clauses [2] and is used as a base for many policy languages, e.g., [3], [4], [5]. Policies in our definition as formulae with one free variable can be expressed as monoid Datalog queries. Compliance checks can be solved via query evaluation, which is decidable. As we will discuss in Section V-B, also all other required operations are decidable for Datalog.

As an example, we formalise the policy BY-NC restricting data usages to trigger an attribution of the original creator (see Section II):

$$\text{BY-NC}(x) \leftarrow \mathsf{Usage}(x) \wedge \mathsf{triggers}(x,a) \wedge \mathsf{Attribution}(a) \wedge \\ \mathsf{hasPurpose}(x,r) \wedge \mathsf{NonCommercial}(r).$$

In Datalog, the variables a and r are existentially quantified, which means that the right-hand side of the rule is a FOL formula with one free variable (x) and thus defining $\phi_{\text{BY-NC}}$.

IV. Distinguishing Obligations and Violations

In situations, where a data usage is classified non-compliant to the used artefact's policy, we have to distinguish between policy violations and not yet fulfilled obligations. Obligations are temporary violations of a policy, which will be fixed after a certain amount of time to reach compliance. Consider for example the obligation to attribute the original creator of an artefact when it is used: using the artefact is classified as non-compliant but only temporarily until the attribution is given and thus compliance reached. If usage is restricted to non-commercial purposes and a usage is classified as non-compliant because it has a commercial purpose, the violation is not temporary and thus there is not an obligation required, but the usage should be prevented. In the following, we present an approach to distinguish violations and obligations for usages classified as non-compliant to a policy.

Consider a data usage described by the theory T, where a policy subject s is found non-compliant to a policy p defined by $\phi_p[x]$. The solution is structured along the following steps:

1) finding out why s is non-compliant;
2) if s can be made compliant by adding new facts, identify the required facts;
3) identify obligations in the facts;
4) checking whether obligation handling makes the usage compliant;
5) if compliance is given, schedule the obligations with the corresponding handlers.

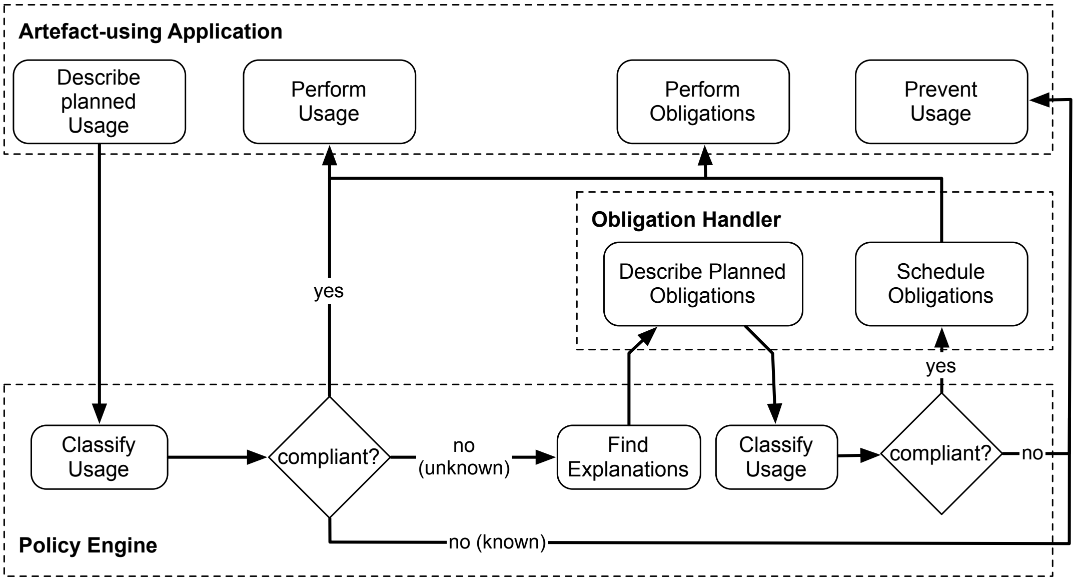


Figure 2. Architecture of a Policy-aware System for Data Usage and Automated Obligation Handling

Step 1: Finding out why s is non-compliant:
We consider s to be non-compliant with p, if we cannot infer that s makes ϕ_p true, i.e., $T \not\models \phi_p[s]$. One reason why we cannot infer $\phi_p[s]$ can be that it contradicts T, i.e., $T \models \neg\phi_p[s]$. In case of a contradiction, we cannot establish $\phi_p[s]$ by adding new facts (e.g., from describing the fulfilment of an obligation), because of the monotonicity of FOL. As such contradictions cannot be fixed by obligation handling, we only proceed if $T \not\models \phi_p[s]$ and $T \not\models \neg\phi_p[s]$.

As an example, consider the following theories describing data usages:

$T_1: \mathsf{Usage(u1)} \wedge \mathsf{hasPurpose(u1, r1)} \wedge \mathsf{NonCommercial(r1)}.$

$T_2: \mathsf{Usage(u2)} \wedge \mathsf{triggers(u2, a2)} \wedge \mathsf{Attribution(a2)}.$

$T_3: \mathsf{Usage(u3)} \wedge \mathsf{hasPurpose(u3, r3)} \wedge \mathsf{Commercial(r3)}.$

All three theories T_1, T_2, T_3 do *not* model a usage compliant to the policy BY-NC. However, $T_3 \models \neg\phi_{\text{BY-NC}}[\mathsf{u3}]$ and thus will be disregarded in further examples.

Step 2: Identify suitable theories to add:
Next, we search for a set $\mathcal{E}$ of theories that make s compliant to p. The search naturally translates into a problem that can be solved by abductive reasoning. The term of abductive reasoning goes back to Peirce [6] and refers to finding an explaining hypothesis for a circumstance. In other words, for a given observation b find an explanation a from which b can be logically inferred. In this sense, abduction is the reverse of of deduction, where b is found for a given a. Abductive reasoning was applied to formal logics and several algorithms were given for various logic formalisms (e.g., [7], [8], [9], [10]). In the following, we formally define our understanding of abductive reasoning for FOL. Given a theory T and a set F of atomic facts (formulae of the form $p(c_1, \ldots, c_n)$, where $p \in P$ is a predicate of arity n, and each $c_i \in C$ is a constant), find an explanation E, such that F can be inferred from T and E, or more formally: $T \cup E \models F$. Additionally, we require that there exists an interpretation for $T \cup E$, i.e., $T \cup E$ is consistent. For sake of simpler notation, we also apply abduction to find an explanation for a sentence $\phi[c]$, where $\phi[x]$ is a formula with the only free variable x. This can be realised by introducing a fresh unary predicate p' and the axiom $\forall x.p'(x) \leftrightarrow \phi[x]$; then abduction can be applied to finding an explanation for the atomic fact $p'(c)$. Applying abduction to our problem of finding suitable theories for making s compliant to p, we search a set $\mathcal{E}$, such that: $\forall E \in \mathcal{E}. T \cup E \models \phi_p[s]$. We require that every explanation E is minimal in the sense that there is no other explanation E' which entails E and there is no subtheory of E, which is also an explanation:

$$\forall E \in \mathcal{E}.\ \nexists E' \in \mathcal{E}. E \neq E' \wedge T \cup E' \models T \cup E.$$

$$\forall E \in \mathcal{E}.\ \nexists E'. E' \subseteq E \wedge T \cup E' \models \phi_p[s].$$

The set of explanations can still be of infinite size, e.g., because of transitive predicates, and the minimality conditions might not always be desired [7]. We leave the exact definition of the explanations selected for $\mathcal{E}$ open to be specified for concrete applications. Similarly, there maybe a system-specific preference order on the explanations, therefore we describe the following steps for a single explanation $E \in \mathcal{E}$.

Continuing the previous examples, we choose the following explanation E_1, E_2 such that $T_1 \cup E_1 \models \phi_{\text{BY-NC}}[\mathsf{u1}]$ and

$T_2 \cup E_2 \models \phi_{\text{BY-NC}}[\mathsf{u2}]$:

$$E_1: \mathsf{triggers}(\mathsf{u1}, \mathsf{a1}) \wedge \mathsf{Attribution}(\mathsf{a1}).$$
$$E_2: \mathsf{hasPurpose}(\mathsf{u2}, \mathsf{r2}) \wedge \mathsf{NonCommercial}(\mathsf{r2}).$$

Step 3: Identification of obligations:
An explanation E contains facts that would make s compliant to p. Not all of the facts in E however can be fulfilled by adding the description of an obligation, but could only be the result of complying to an unfulfilled condition (see example below). Depending on the specific application, we thus define a set $\mathcal{O}$ of obligations, and for every obligation $o \in \mathcal{O}$ a query $q_o(p_1, \ldots, p_n)$ and an obligation handler h_o. The query q_o defines, which kind of required facts can be handled by the corresponding obligation handler h_o. In our example, we define one obligation $o1$ with a handler h_{o1} that can automatically add attributions to data usages. The corresponding query q_{o1} is defined as:

$$q_{o1}(x, a) \leftarrow \mathsf{triggers}(x, a) \wedge \mathsf{Attribution}(a).$$

The bindings for the query are passed to the obligation handler h_o, which will return a FOL theory T' that describes the planned fulfilment of the obligations identified by the bindings. In our example, for E_1 the query q_{o1} gives the binding $\{x \mapsto \mathsf{u1}, a \mapsto \mathsf{a1}\}$, for which the handler h_{o1} plans to create an attribution action, described by the returned theory:

$$T'_1: \mathsf{triggers}(\mathsf{u1}, \mathsf{a1}') \wedge \mathsf{Attribution}(\mathsf{a1}').$$

For the explanation E_2, the query q_{o1} gives no bindings, and thus the obligation handler only returns the empty theory T'_2.

Step 4: Checking if obligation handling leads to compliance:
After getting the descriptions of the planned obligation fulfilments, we want to ensure that fulfilling them is sufficient to make s compliant. For this we check whether $T \cup T' \models \phi_p[s]$. If this is the case, we can proceed to the next step and schedule the planned obligation fulfilments. Otherwise, we found out that s is not only a temporary violation, but should be prevented completely. In our example, we see that $T_1 \cup T'_1 \models \phi_{\text{BY-NC}}[\mathsf{u1}]$, but $T_2 \cup T'_2 \not\models \phi_{\text{BY-NC}}[\mathsf{u2}]$. Thus, we prevent $\mathsf{u2}$ from execution, but allow $\mathsf{u1}$ and tell the obligation handler h_{o1} to schedule the attribution $\mathsf{a1}'$ (**Step 5: Obligation handling**).

A system architecture realizing the complete process of Steps 1 to 5 is visualised in Figure 2.

In order, to ensure that every obligation can be unambiguously assigned to an obligation handler, one can require that the q_o queries define pairwise disjoint sets for different obligations. Another, weaker, requirement would be that no obligation definition is subsumed by an other definition. In some systems, however, it may also be practical to pose no such requirements and have an obligation subsuming all other obligations, which has an handler that logs all obligation instances.

V. Implementation and Application

In this section, we describe how the proposed concepts can be used for realising the use case of data usage restrictions. We then argue, how two popular policy formalisms (Datalog and OWL) can be used with our approach, by describing how the required operations can be realised with standard reasoner methods. Finally, we briefly describe how we implemented the approach for Datalog-based policies.

A. Realisation of Use Case

We already discussed the Creative Commons NonCommercial, Attribution policy and its application to three different usages as a running example in the explanation of our approach. The other two policies are given in the following:

- Creative Commons ShareAlike (abbreviated SA): derived artefact should have the same policy:

  $$\mathsf{SA}(x) \leftarrow \mathsf{Derivation}(x) \wedge \mathsf{wasGenBy}(a, x) \wedge \mathsf{hasPolicy}(a, \mathsf{SA}).$$

  Assigning an allowed target policy for a generated artefact, can be done automatically by an obligation handler h_{o2} taking the bindings of the following obligation query: $q_{o2}(a, p) \leftarrow \mathsf{hasPolicy}(a, p)$. If the obligation handler receives bindings that would assign incompatible policies to an artefact, the obligation handler returns an empty theory back, meaning that it cannot fulfil the obligations. Otherwise, it returns a theory describing that a compatible policy is assigned to the artefact, which is scheduled in case that the obligation descriptions make the usage compliant.
- Electronic health record policy (abbreviated EHR): doctors can use the artefact and store it for one year:

  $$\begin{aligned}\mathsf{EHR}(x) \leftarrow &\Big(\mathsf{Usage}(x) \wedge \mathsf{performedBy}(x, a) \wedge \mathsf{Doctor}(a)\Big) \vee \\ &\Big(\mathsf{Storage}(x) \wedge \mathsf{performedBy}(x, a) \wedge \mathsf{Doctor}(a) \wedge \\ &\quad \mathsf{triggers}(x, d) \wedge \mathsf{Deletion}(d) \wedge \\ &\quad \mathsf{performedAt}(d, t) \wedge t \leq \text{now() + 1y}\Big).\end{aligned}$$

  To a storage action, which is classified as noncompliant, at least one of the following applies: (i) it is not performed by a doctor, or (ii) there is no deletion scheduled. The former cannot be handled as an obligation: allowing only doctors access to the health record is a hard constraint, which cannot even temporarily be violated. In contrast, an automated deletion can be scheduled by an obligation handler in the future, making the storage action compliant. The corresponding obligation query is given as
  $q_{o3}(d, t) \leftarrow \mathsf{Deletion}(d) \wedge \mathsf{performedAt}(d, t)$.

B. Applicability to Concrete Policy Formalisms

We used a Datalog-based policy formalism in this work, as it is a popular choice for policy languages and has desirable computational properties: compliance checks and obligation identification can be solved via query evaluation, which is decidable. Checking whether an obligation query is subsumed by another query is a query containment problem, which can be reduced to query evaluation [11]. Testing whether obligation queries are disjoint is equivalent to testing disjointness of database views and queries, for which algorithms exist [12]. Finally, there exist numerous approaches to abductive reasoning that can be applied to Datalog, e.g., [8], [10].

Other fragments of FOL, for which practical tools exist, are represented by Description Logics [13], namely the Web Ontology Language (OWL) and its profiles [14]. OWL is also used for policy languages (e.g., [15], [16]) and defines concepts, which correspond to formulae with one free variable, and thus is compatible to our approach. Standard inference tasks for OWL reasoners cover almost all required tasks for our procedure presented in this paper: instance classification (for compliance checks and obligation identification), class subsumption (for checking subsumption of obligation queries) and class disjointness checks (for testing disjointness of obligation queries). Missing is only abduction, which is not regarded as a standard task, but solved by several approaches, e.g., [9].

C. Implementation

We developed a prototypical implementation of our approach for Datalog-based policy languages, as described in this paper. The prototype uses the DLV System [17] for compliance classification and a custom obligation handling system based on the abductive reasoning engine HYPROLOG [18] running on SWI-Prolog [19]. The prototype is not optimised, but is able to classify simple examples based on our use case in less than 0.2 seconds and find and handle obligations in less than 1 second on a 2.4 GHz standard laptop computer. The conducted measurements show that an integration into a fast *design, compliance check, modification life cycle* is possible. More extensive performance measurements will be conducted in future work, when the policy and obligation engine is integrated into a concrete policy-aware system for exposing compositions of data artefacts on the Web.

VI. Related Work

As noted before, the term of abductive reasoning goes back to Peirce [6] and many technical solutions for different logic formalisms were developed, e.g., [7], [8], [10], [9]. Related to our task to find out the reasons for a policy non-compliance are so-called *why not*, respectively *how to* questions [20]. Becker and Nanz explicitly mention the use of abductive reasoning in policy systems to determine what is missing to reach compliance [8]. Not targeted at policies but at formal knowledge systems in general is the work of Chalupsky and Ross for answering *why not* queries, i.e., giving reasons why some desired inference does not hold [21]. The applications of explanations and abduction to policies have in common that they aim at helping the user to reach compliance. Our goal is to automatically identify obligations and pass them to an obligation handler. Not all missing pieces described by an explanation can just be regarded as obligations, but could also be violations of the policy. Finding out, which pieces are obligations and whether they cover the full explanation is a non-trivial task for a policy-based system, for which we presented to the best of our knowledge, the first solution.

Xu and Fong present a policy language with obligations [22], for which they list a set of requirements taken from surveying obligation policy languages in the literature, including [23], [24], [25], [26], [27]. In contrast to the languages analysed by Xu and Fong and the language they propose, there are no special logic operators for representing obligations in our approach. Instead, domain- and application-specific types of obligations can be defined. We model only desired goal states, i.e., the states compliant to a policy, and leave computation of what has to be done to reach compliance (including the fulfilment of obligations) to the policy system. This is in contrast to the other approaches, which specify the actions that have to be performed directly using the obligation operators. In the following, we describe how the requirements identified by Xu and Fong [22] are handled by our approach:

- *Trigger* and *obligation*: define under which conditions the obligation is applicable, and what the obligation is. In our approach, both are described in one logical formula specifying the desired and compliant goal states.
- *Temporal constraint*: specifies the time span in which an obligation should be fulfilled. In our approach, this can be modelled, if needed, as part of the domain ontology. Depending on the application, different models of time spans can be employed, e.g., (i) attribution must be given at the same time as usage, or (ii) deletion of artefact must take place latest one year after it was stored.
- *Penalty* or *reward*: what happens if the obligation is violated (*penalty*), respectively fulfilled (*reward*). A penalty can just be modelled as another possibility to reach compliance, namely by executing the protected actions and fulfilling the penalty instead of the obligation. A reward is simply a more relaxed policy, i.e., allowing more actions if the obligation is also fulfilled.

VII. Conclusions

We presented a novel approach to formalise obligations and other compliance conditions in a uniform way. The approach enables goal-based policies on a high conceptual level without the need for users to learn special operators in the policy language. Instead definitions of obligations can be specified using domain- and application-specific vocabularies that are defined and understood by the users. Explanations about what a policy subject lacks to compliance are found by abductive reasoning. We presented a novel method to check whether an explanation is fully covered by obligations and to identify the obligations.

For future work, we plan to integrate the implemented method in a concrete application, realising the automated handling of obligations when using data with restricted usages to create new services and data sources.

Acknowledgements: This work was supported by the EU project PlanetData (FP7 - 257641).

REFERENCES

[1] J. O. Kephart and W. E. Walsh, "An Artificial Intelligence Perspective on Autonomic Computing Policies," in *Proceedings of the Fifth IEEE International Workshop on Policies for Distributed Systems and Networks*, 2004, pp. 3–12.

[2] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*. Addison Wesley, 1994.

[3] P. A. Bonatti, J. L. De Coi, D. Olmedilla, and L. Sauro, "A Rule-based Trust Negotiation System," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1507–1520, 2010.

[4] C. Ringelstein and S. Staab, "PAPEL: A language and model for provenance-aware policy definition and execution," in *Proceedings of 8th International Conference on Business Process Management (BPM)*, ser. LNCS, R. Hull, J. Mendling, and S. Tai, Eds., vol. 6336. Springer, 2010, pp. 195–210.

[5] M. Y. Becker, C. Fournet, and A. D. Gordon, "SecPAL: Design and Semantics of a Decentralized Authorization Language," *Journal of Computer Security (JCS)*, vol. 18, pp. 597–643, 2010.

[6] C. S. Peirce, "Abduction and Induction," in *Philosophical writings of Peirce*. Dover Publications, 1955.

[7] D. Poole, "Explanation and prediction: an architecture for default and abductive reasoning," *Computational Intelligence*, vol. 5, pp. 97–110, May 1989.

[8] M. Y. Becker and S. Nanz, "The Role of Abduction in Declarative Authorization Policies," in *Proceedings of the 10th International Symposium on Practical Aspects of Declarative Languages (PADL)*, ser. LNCS, P. Hudak and D. Warren, Eds., vol. 4902. Springer, 2008, pp. 84–99.

[9] S. Klarman, U. Endriss, and S. Schlobach, "Abox abduction in the description logic $\mathcal{ALC}$," *Journal of Automated Reasoning*, vol. 46, pp. 43–80, 2011.

[10] U. Endriss, P. Mancarella, F. Sadri, G. Terreni, and F. Toni, "The CIFF Proof Procedure for Abductive Logic Programming with Constraints," in *European Conference on Logics in Artificial Intelligence*, ser. LNCS, J. Alferes and J. Leite, Eds. Springer, 2004, vol. 3229, pp. 31–43.

[11] S. Abiteboul and O. M. Duschka, "Complexity of answering queries using materialized views," in *Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (PODS)*. ACM, 1998, pp. 254–263.

[12] M. W. Vincent, M. Mohania, and M. Iwaihara, "Detecting privacy violations in database publishing using disjoint queries," in *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology (EDBT)*. ACM, 2009, pp. 252–262.

[13] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press, 2003.

[14] W3C OWL Working Group, *OWL 2 Web Ontology Language: Document Overview*. W3C Recommendation, 27 October 2009, available at http://www.w3.org/TR/owl2-overview/.

[15] M. Krötzsch and S. Speiser, "ShareAlike your data: Self-referential usage policies for the Semantic Web," in *Proceedings of the 10th International Semantic Web Conference (ISWC)*, ser. LNCS, vol. 7032. Springer, 2011, pp. 354–369.

[16] A. Uszok, J. Bradshaw, R. Jeffers, N. Suri, P. Hayes, M. Breedy, L. Bunch, M. Johnson, S. Kulkarni, and J. Lott, "KAoS Policy and Domain Services: Toward a Description-Logic Approach to Policy Representation, Deconfliction, and Enforcement," in *Proceedings of the 4th IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY)*, 2003, pp. 93–96.

[17] "DLV," (Software) http://www.dlvsystem.com/dlvsystem/index.php/DLV, accessed March 14th 2012.

[18] "HYPROLOG," (Software) http://akira.ruc.dk/~henning/hyprolog/, accessed March 14th 2012.

[19] "SWI-Prolog," (Software) http://www.swi-prolog.org/, accessed March 14th 2012.

[20] P. A. Bonatti, D. Olmedilla, and J. Peer, "Advanced Policy Explanations on the Web," in *Proceedings of the 17th European Conference on Artificial Intelligence (ECAI)*. IOS Press, 2006, pp. 200–204.

[21] H. Chalupsky and T. Russ, "Whynot: Debugging failed queries in large knowledge bases," in *Proceedings of the 14th conference on Innovative Applications of Artificial Intelligence (IAAI)*. AAAI Press, 2002, pp. 870–877.

[22] C. Xu and P. W. L. Fong, "The Specification and Compilation of Obligation Policies for Program Monitoring," Department of Computer Science, University of Calgary, Canada, Tech. Rep. 2011-996-08, April 2011, available at http://pages.cpsc.ucalgary.ca/~pwlfong/Pub/UC-CPSC-TR-2011-996-08.pdf.

[23] N. H. Minsky and A. D. Lockman, "Ensuring integrity by adding obligations to privileges," in *Proceedings of the 8th International Conference on Software Engineering (ICSE)*. IEEE Computer Society Press, 1985, pp. 92–102.

[24] N. Damianou, N. Dulay, E. Lupu, and M. Sloman, "The Ponder Policy Specification Language," in *Proceedings of the Workshop on Policies for Distributed Systems and Networks (POLICY)*, ser. LNCS, vol. 1995. Springer, 2001, pp. 18–38.

[25] P. Gama and P. Ferreira, "Obligation Policies: An Enforcement Platform," in *IEEE Workshop on Policies for Distributed Systems and Networks (POLICY)*, 2005, pp. 203–212.

[26] K. Irwin, T. Yu, and W. H. Winsborough, "On the modeling and analysis of obligations," in *ACM Conference on Computer and Communications Security (CCS)*, 2006, pp. 134–143.

[27] M. Hilty, A. Pretschner, D. Basin, C. Schaefer, and T. Walter, "A policy language for distributed usage control," in *European Symposium on Research in Computer Security (ESORICS)*, 2007, pp. 531–546.

Towards a Reuse-oriented Security Engineering for Web-based Applications and Services

Aleksander Dikanski, Sebastian Abeck
Research Group Cooperation & Management (C&M)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
{ a.dikanski, abeck }@kit.edu

***Abstract*—Security should be considered throughout a software development process to develop secure applications. This security engineering effort is restricted due to the complexity and diffusion of todays security knowledge. Approaches, such as misuse cases for threat specification and patterns for security functionality modeling, try to use and integrate security into software development, but their combined use is still difficult. In this paper a framework for developing secure software systems is presented, which aims at incorporating and unifying existing security engineering approaches by applying well-established reuse-oriented software development paradigms, such as service-orientation. The security-related activities and reusable artifacts of important development phases are discussed and the mapping of artifacts between different development phases is presented.**

Keywords-security engineering; software development; security patterns; service-orientation

I. Introduction

The increasing number of attacks on software systems makes it more important than ever to develop secure software systems. Especially web-based applications and services are faced with numerous threats due to their public access. But, the prevailing custom of including security functionality after the functional development is infeasible, is not fulfilling the actual security needs. Security engineering aims for a consecutive secure software develop-ment by introducing methods, tools, and activities into a software development process [1].

Such an integration has not yet been achieved completely as the amount of security knowledge, including theoretic models, technologies and standards, developed until now is complex, often diffused, and seldom structured enough to be used in a software development process. Opposed to this, security can usually be considered reusable across heterogeneous functional domains, e.g., access control models such as role-based access control (RBAC, [2]) can be used in different domains. Yet, so far structured means for reuse of security functionality are not successfully employed. Existing approaches contribute mainly to specific develo-pment phases. Yet, while each of these approaches is beneficial in its intentions, they are hard to integrate.

In this work, an early version of a framework is presented, which aims at structuring existing security knowledge in a reusable fashion and providing decision support to integrate existing security engineering approaches and methodologies more concisely. The concepts of modern reuse-oriented paradigms, such as service-orientation, software product lines (SPL) as well as model-driven software development (MDSD), are facilitated in our approach.

The goal is to present developers a structured tool set, to ease the coherent integration of security aspects into each phase and across phases. Thus an increased quality of the security functionality is achieved.The framework comprises security requirements analysis templates and security pattern languages. The former can be instantiated to analyze the security needs of an application in a deterministic way, while the latter can be used to choose appropriate security solutions and iteratively refine them.

In the next section, approaches relevant for our work will be discussed. In Section 3, the contribution of our approach will be described. We further present two projects which lead to the development of our framework in Section 4. A conclusion closes the body of this paper.

II. Related Work

Reuse in security engineering processes is discussed in several approaches. The SECTET-framework [3] provides a service-oriented security engineering approach for authorization in inter-organizational workflows, but concentrates mostly on web-service based architectures. The Secure-Change-Project aims at a change-driven security engineering approach, in which security requirements are specified, which evolve throughout the lifetime of software [4]. While the focus of this project is change of security requirements and security design, our focus lies on presenting feasible choices and decision support for them to increase quality of security functionality.

Threat and risk analysis techniques for analyzing and specifying security requirements include STRIDE [5], attack trees [6], and misuse cases [7]. We are aiming at providing deterministic threat descriptions at an appropriate abstraction level and link them to appropriate security requirement specifications to complement theses techniques, as this is were each of them fails short and is thus difficult to apply.

Security patterns are a popular and widely accepted method for modeling technology-independent security functions [8]. Security pattern languages are utilized to describe the connections between multiple patterns and their combined usage [9]. But, alternative solutions are not considered by existing languages. So far, only SPL approaches consider such variations [10]. We aim to enhance

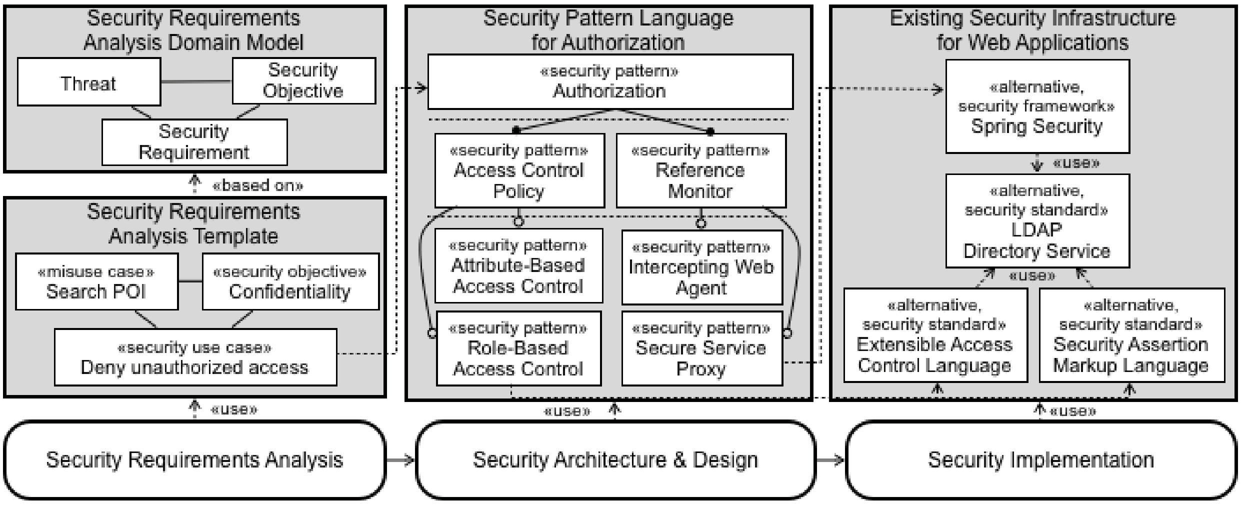

Security Requirements Analysis Domain Model
Threat
Security Objective
Security Requirement
«based on»
Security Requirements Analysis Template
«misuse case» Search POI
«security objective» Confidentiality
«security use case» Deny unauthorized access
«use»
Security Pattern Language for Authorization
«security pattern» Authorization
«security pattern» Access Control Policy
«security pattern» Reference Monitor
«security pattern» Attribute-Based Access Control
«security pattern» Intercepting Web Agent
«security pattern» Role-Based Access Control
«security pattern» Secure Service Proxy
«use»
Existing Security Infrastructure for Web Applications
«alternative, security framework» Spring Security
«use»
«alternative, security standard» LDAP Directory Service
«use»
«alternative, security standard» Extensible Access Control Language
«alternative, security standard» Security Assertion Markup Language
«use»
Security Requirements Analysis
Security Architecture & Design
Security Implementation

Figure 1. Overview over the security engineering framework

security pattern approaches by explicitly showing alternative pattern solutions to security problems.

Model-driven security applies to methods of model-driven software development to the security domain. Secure-UML [11], UMLSec [12], and the work of Emig et al. [13] are among the most prominent approaches in this field. They do not consider existing security infrastructures in a service-oriented way as we intend to do. Also, they also do not provide chocies between alternative security patterns.

III. Security Engineering Framework

The framework presented in the following complement and unifyies existing approaches in the security engineering field by providing reusable security-related development artifacts and a decision support for them.

Reuse is at the core of many well-established software engineering paradigms, which aim at managing complex software systems development, such as service-orientation, SPL and MDSD. A security engineering methodology based on the reuse of existing security knowledge will lead to an increased efficiency in the development of secure software and to an improved quality of the security functionality.

An important goal for our approach is to be development process agnostic, i.e., the artifacts contributed by our framework should be independent from specific software development processes and instead be applicable in different methodologies and paradigms.

We further aim for decision support and guidance in using security knowledge. The security domain comprises a large knowledge base, including, e.g., security standards and technologies as well as security models, principles and policies. A structured approach is needed for applying this knowledge in a development process. The focus lies on supporting a decision process by pointing out alternative solution to security problems.

Currently, the framework is limited to security within web application and service development, thereby neglecting lower levels of security measures such as web server, operating system, and network, even though this is considered bad security practice. Yet, we do not rule out the applicability of our approach to these levels.

The next sections will present the core elements of our framework and their intented function. The focus, thereby, lies upon the first three phases, i.e., requirements analysis, design and implementation phase. Testing and operation are important phases in the development of secure applications as well, but we exclude them here for brevity reasons.

A. Reusable Security Requirements Templates

Similar to functional requirements elicitation, security requirements need to be analyzed and specifed as well to determine application security needs. Difficulties in this phase concern the appropriate abstraction level and the format of the security requirements specification. Often they are specified by proposing security functionality, instead of constraints to the functionality [14][15].

According to our goals, our approach strives for contributing reusable *security requirements analysis templates (SecRAT)* to this phase. The template's core is based on the relationships between threats, which violate security objectives, security requirements, which are capable of mitigating the threats and implement security objectives. These entities and their relationships form a basic *security requirements analysis domain model (SecADM)*, giving structure to the templates. The templates will further be categorized into *domain-independent SecRATs*, applicable to multiple functional domains, and *domain-specific SecRATs*, describing security requirements and threats specific to a functional domain. This allows for a more focused and structured approach to requirements analysis.

Reuse of security knowledge is thereby achieved by documenting existing threat knowledge and explicitly linking it to appropriate security requirements and objectives. Therefore, if a threat is determined to be applicable in an application development process, the appropriate template can be instantiated, directly leading to related security requirements as well as objectives and vice versa.

The templates are thereby independent of any approach for analyzing and specifying security requirements such as those mentioned in Section 2. Instead they can be used as a structured decision support tool to determine necessary security requirements as well as a common specification format for any such process and modeling tools. Further, each SecRAT is linked to an abstract security functions, thereby supporting the transfer between requirements engineering and design phase.

B. *Security Pattern Language and Variability Model*

The goal of the design phase is to implement security requirements using appropriate security functions. They are firstly specified at a coarse-grained, technology-independent architecture-level design and iteratively refined to an fine-grained, implementation-level design. These functions form the security architecture of one or more applications and thus need to be integrated into the overall architecture [16].

For the iterative refinement process, a concise *security pattern language (SecPAL)* for each security function design is proposed, which builds upon and complements previous approaches. It enables the use and combination of multiple security patterns, each of which relates to and implements a certain security requirement. Thus a decision support is offered, in that only compatible patterns are connected in the pattern language. Yet, opposed to previous approaches, the focus of the SecPALs lies on iterative refinement.

At each iterative refinement the design is not always obvious. In fact, a choice between several design options can be made. For example, to implement access control, several alternatives exist, including role-based (RBAC, [2]) and attribute-based access control [17], each of which might be more suitable depending on application context.

To provide an overview over viable alternative security patterns applicable to specific security problems, the SecPAL is complemented by a *security pattern variability model (SecPVM)*. In each iterative refinement of a pattern, the variability model can be applied to select an appropriate variant for a pattern, if necessary. Currently, feature models [18], a common tool to model commonality and variability in SPL development, are feasible candidates to describe security pattern variants including mandatory, optional and exclusion relationships.

C. *Service-Oriented Security Design*

The combination of SecRAT, SecPAL, and SecPVM is intended to support the development of new or the extension of existing security functionality for software systems. But they can also be used to support secure development projects, which need to be integrated into an existing security infrastructure, e.g., in an enterprise environment. In this context, we build upon our previous efforts [16][19] by applying the service-orientation paradigm, i.e., the reuse and restructuring of existing software systems to satisfy business needs, to the design phase of security engineering as well.

Reusing existing services narrows security design decisions. When developing an application for an IT infrastructure in which, e.g., RBAC is the standard access control policy model, a decision about the policy model to use for access control in the newly developed application is already determined.

In order to achieve the benefits of using security services in the design models, an abstraction of the implemented services to a technology-independent level is required, displaying appropriate views of the complete security architecture to developers [20].

We currently employ a manual approach, in which required abstractions are provided by security experts once for each utilized product, as we have done in previous work [19]. The SecPAL can in this case be used as guidance to identify fine-grained patterns within existing security frameworks and products. SecPVM can be used to identify and document alternative implementations offered by the security framework or product. By following the language paths in reverse direction, a relationship to more abstract, coarse-grained patterns can be established.

D. *Standards- and Pattern-based Model-Driven Security*

Despite the reuse of existing functionality it is inevitable that certain artifacts need to be developed as part of the security engineering approach, even though our goal is to reduce the number of such artifacts to allow for an efficient development. In this context, we continue our previous efforts on model-driven security [13], but are more focussed on integrating it into a security engineering approach using security technology standards and patterns to automatically generate necessary artifacts.

While implementing security functionality, employing security technology standards offers product independence and interoperability. Yet, applying standards without in depth knowledge is difficult, as they include a large degree of flexibility. A very good example for this is the slowly progressing adoption of XML-based security standards, developed mainly for web service-based applications [21][22]. Note that the same can be argued for security frameworks and products.

As a benefit of the security pattern identification and specification using the SecPAL described in the previous section, specific guidelines and templates on how standards are to be utilized to implement a certain security pattern. As such, we are able to provide a security platform description, which is used as a automatically generate relevant artifacts from design models specified using SecPAL.

IV. Motivating Case Study Scenario

We are currently applying, refining, and evaluating our approach by applying it in the development of two real-world projects, requiring security functionality.

A. *Case Study Description*

The KITCampusGuide (KCG) is a web-based and service-oriented geographic information system (GIS). It supports employees, students and guest of the Karlsruhe Institute of Technology (KIT) with their daily campus activities. Its basic functionality allows the user to search for points of interest (POI), such as buildings, rooms or offices, and display results on a campus map.

This functionality will be extended as a proof-of-concept for the european project OpenIoT by enabling students to search for available workplaces on the campus. This functionality will be implemented using smart objects. These virtual or physical objects, such as rooms, are active participants in the information systems and can be remotely queried and their state modified using sensor and actor technology.

Very early it became clear, that security aspects needed to be implemented in the KCG application as the KIT is restricted by legislative regulations, so that the privacy and anonymity of the users, as well as integrity and confidentiality of the processed data need to be assured, due to the location sensitive nature of the application.

As the application was targeted to be integrated into the overall KIT IT-infrastructure, utilization of the existing security infrastructure was required. Therefore, the capabilities of the provided security services need to be analyzed, so that the analyzed security requirements can later be mapped upon them.

B. *Preliminary Results*

We are currently in the progress of formulating an initial set of reusable SecRATs for web-based GIS and IoT applications based on our experiences in developing the KCG application as well as best practice security requirements found in literature. These SecRATs are extensions of a domain-independent SecRAT, which is currently developed as well. From these, a first draft of the SecADM will be developed. For the documentation we are currently using textual templates, but are evaluating more formal methods.

We are further using existing works on security patterns to formulate appropriate SecPALs for different security solutions such as authentication and authorization. Not many such patterns are available in the field of IoT-applications, which is why we will document new security solutions as well. In doing so, we are evaluating and formalizing alternative design decisions into an appropriate SecPVM using feature models.

We plan to validate our approach by applying it to other projects from the same domain to test the integrity of the developed artifacts. We further plan to adopt our approach to different domains including cloud based applications. To evaluate the increased quality of the implemented security functionality empirical studies will be performed.

V. Conclusion and Outlook

In this paper, a security engineering framework was outlined, which focuses on the structured reuse of existing security knowledge by providing analysis templates and pattern languages for security to increase quality of the implemented security functionality. We argued the benefits of security engineering and pointed out short-comings of existing approaches due to missing integration and combination. We identified that a structured approach need to be based on reuse of existing security knowledge and a decision support system in order to be feasible in the field. We presented an outline for several contributions to the different phases of a development process based on and complementing existing approaches, which we think will be beneficial to more efficient security engineering. In future works, we will flesh out the details of the contributions. We further presented current projects, which are used as cases studies to demonstrate the practicability of our approach.

References

[1] R. J. Anderson, Security Engineering. 2nd ed., Indianapolis, Ind.: Wiley, 2008, p. 1040.

[2] D. Ferraiolo, R. Sandhu, S. Gavrila, and D. Kuhn, “Proposed NIST Standard for Role-Based Access Control,” ACM Transactions on Information and System Security (TISSEC), vol. 4, no. 3, pp. 224–274, 2001.

[3] M. Hafner and R. Breu, Security Engineering for Service-Oriented Architectures. Heidelberg: Springer, 2008.

[4] R. Scandariato and F. Massacci, “SecureChange: security engineering for lifelong evolvable systems,” in ISoLA'10: Proceedings of the 4th international conference on Leveraging applications of formal methods, verification, and validation, 2010, vol. Part II , Volume Part II, pp. 9–12.

[5] D. Verdon and G. McGraw, “Risk analysis in software design,” IEEE Security & Privacy, vol. 2, no. 4, pp. 79–84, 2004.

[6] B. Schneier, Secret & Lies, 1st ed., Weinheim: dpunkt.verlag, 2001, p. 408.

[7] G. Sindre and A. L. Opdahl, “Eliciting security requirements with misuse cases,” Requirements Engineering, vol. 10, no. 1, pp. 34–44, 2005.

[8] M. Schumacher, E. B. Fernandez, D. Hybertson, F. Buschmann, and P. Sommerlad, Security Patterns. Chichester, England: John Wiley & Sons Ltd, 2005, p. 565.

[9] E. B. Fernandez and R. Pan, “A Pattern Language for Security Models,” Conference on Pattern Languages of Programs, 2001.

[10] T. E. Fægri and S. O. Hallenstein, “A Software Product Line Reference Architecture for Security,” in Software Product Lines, no. 8, Berlin, Heidelberg: Springer, 2006, pp. 276–326.

[11] T. Lodderstedt, D. Basin, and J. Doser, “SecureUML: A UML-Based Modeling Language for Model-Driven Security,” LNCS, vol. 2460, pp. 426–441, 2002.

[12] J. Jürjens, “Model-Based Security Engineering with UML,” in Foundations of Security Analysis and Design III, vol. 3655, no. 2, A. Aldini, R. Gorrieri, and F. Martinelli, Eds. Berlin, Heidelberg: Springer, 2005, pp. 42–77.

[13] C. Emig, S. Kreuzer, S. Abeck, J. Biermann, and H. Klarl, “Model-Driven Development of Access Control Policies for Web Services,” Proceedings of the 9th IASTED International Conference Software Engineering and Applications, vol. 632, pp. 069–165, 2008.

[14] J. D. Moffett, C. B. Haley, and B. Nuseibeh, “Core security requirements artefacts,” Department of Computing, Milton Keynes, 2004/23, 2004.

[15] J. Rushby, “Security requirements specifications: How and what,” Symposium on Requirements Engineering for Information Security (SREIS), vol. 441, 2001.

[16] C. Emig, F. Brandt, S. Kreuzer, and S. Abeck, “Identity as a Service-Towards a Service-Oriented Identity Management Architecture,” LNCS, vol. 4606, pp. 1–8, 2007.

[17] E. Yuan, J. Tong, B. Inc, and V. McLean, “Attributed Based Access Control (ABAC) for Web Services,” 2005 IEEE International Conference on Web Services, 2005.

[18] K. Lee, K. C. Kang, and J. Lee, “Concepts and Guidelines of Feature Modeling for Product Line Software Engineering,” in Software Reuse: Methods, Techniques, and Tools, vol. 2319, no. 5, C. Gacek, Ed. Berlin, Heidelberg: Springer, 2002, pp. 62–77.

[19] A. Dikanski, C. Emig, and S. Abeck, “Integration of a Security Product in Service-oriented Architecture,” in 2009 Third International Conference on Emerging Security Information, Systems and Technologies, Athens, Greece, 2009, pp. 1–7.

[20] A. Dikanski and S. Abeck, “A View-based Approach for Service-Oriented Security Architecture Specification,” in The Sixth International Conference on Internet and Web Applications and Services, St. Maarten, The Netherland Antilles, 2011.

[21] T. Imamura and M. Tatsubori, “Patterns for Securing Web Services Messaging,” OOPSLA Workshop on Web Services and Service Oriented Architecture Best Practice and Patterns, pp. 1-8, 2003.

[22] N. A. Delessy and E. B. Fernandez, “Patterns for the eXtensible Access Control Markup Language,” Proceedings of the 12th Pattern Languages of Programs Conference (PLoP2005), pp. 7–10, 2005.

A Stream-Oriented Community Generation for Integrating TV and Social Network Services

Riho Nakano
Faculty of Policy Management
Keio University
Fujisawa, Kanagawa, Japan
e-mail: s10578rn@sfc.keio.ac.jp

Shuichi Kurabayashi
Faculty of Environment and Information Studies
Keio University
Fujisawa, Kanagawa, Japan
e-mail: kurabaya@sfc.keio.ac.jp

***Abstract*— This paper proposes a stream-oriented community generation that associates real-time TV streams and Web resources by analyzing the communication among users on a social network service (SNS). The aim of this system is to provide a novel media environment for enhanced cross-media communication and discussion by dynamically creating social communities according to the real-time contexts of TV stream. The unique feature of this system is an implicit community analysis mechanism that employs the TV stream as a powerful and well-organized "context-creator" for SNS users. This system extract a group of viewers who have the same or similar interests by integrating the term co-occurrence statistics of SNS messages and their synchronicity to TV. To detect the context-dependent group of users, this system provides a dynamic feature keyword selection mechanism to create a vector space, which is specifically tailored to the TV context. The application scope of this system includes analysis of community-level sentiments in SNS messages associated with a TV program and the analysis of transitions in the sentiments of communities to develop effective advertising strategies.**

Keywords- Community Generation; Social Network Services; TV streams; Cross-media Infrastructure; Sentiment Analysis.

I. INTRODUCTION

With the popularization of Internet-enabled TVs and smartphones, the demand for integrating synchronous TV with asynchronous SNSs has increased because their relationship is complementary [1, 2]. TV motivates viewers to interact with each other by generating common interests among them, and SNS enables this interaction. Integrating TV with interactive services increases its value [3].

Although TVs and SNSs are popular information resources, studies on the implicit community analysis and community creation by integrating these media have been limited. This is because the current Internet TV (ITV) technologies focus on the integration of video streams with textual information retrieved from the Web by developing screens that are capable of displaying Web contents. Therefore, cross-media communication infrastructure is essential to get the integrated sentiment information and its context which are present in a fragmented and closed manner.

In this paper, a stream-oriented community generation is proposed to create a group of viewers, i.e., a community of users, who have common interests by integrating real-time TV streams and SNS messages. "Community Generation", which is used to widen the scope of communication by adopting a dynamic configuration, is the key concept in such communication infrastructure [4]. Community Generation enables viewers to broadcast messages about the current TV program to the appropriate audience. In order to realize community generation, this system extracts an implicit structure of viewers who have the same or similar interests by analyzing their comments about a TV stream. The association between a community of TV viewers and a community of SNS users are established according to the information appearing on both the TV and the SNS. This system automatically extracts a "community of interest" (CoI) structure from the SNS messages and TV program guides. Unlike Google TV that uses a single display for synchronous TV streams and asynchronous SNS messages, this system enables cross-media communication between the two information resources.

The aim of this system is to provide a novel media environment where a TV stream and related messages are exchanged seamlessly according to its context. As shown in Figure 1, this system uses the TV stream as a powerful and well-organized "context-creator" for SNS users. The context-creator affects a vast array of users by posting the same content at the same time over many SNSs. By introducing TV as a context-creator into SNS community analysis, the system recognizes the topic in the SNS messages by leveraging the powerful context created by the TV.

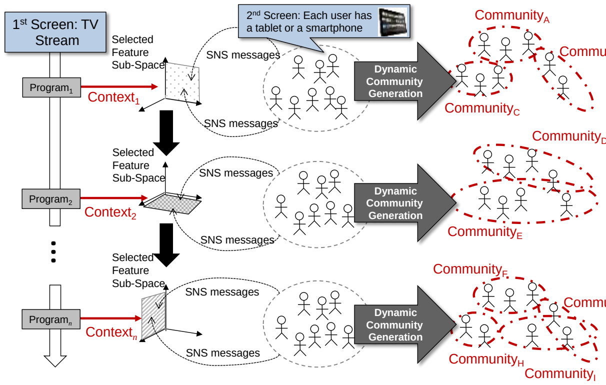


Figure 1. Fundamental Concept of Stream-Oriented Community Generation.

The most important advantages of this system are recognition of overlapping structures among communities and tracking viewers' transitions among multiple communities by detecting the sentiments expressed in SNS messages related to a TV stream. This tracking mechanism

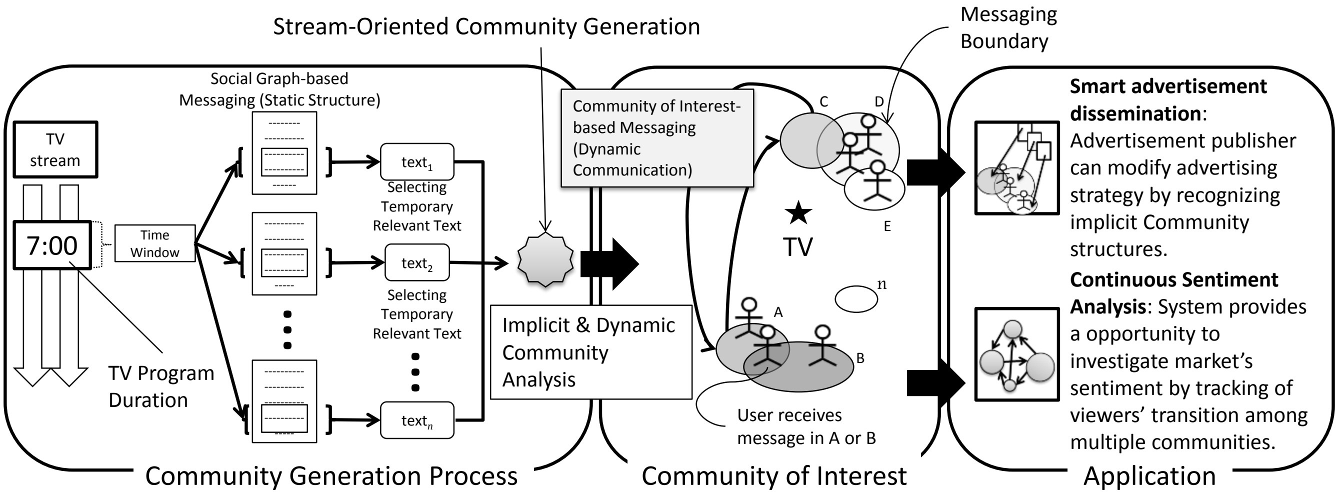

Stream-Oriented Community Generation
Messaging Boundary
Social Graph-based Messaging (Static Structure)
TV stream
7:00
Time Window
TV Program Duration
text$_1$
Selecting Temporary Relevant Text
text$_2$
Selecting Temporary Relevant Text
text$_n$
Community of Interest-based Messaging (Dynamic Communication)
Implicit & Dynamic Community Analysis
C
D
E
TV
n
A
B
User receives message in A or B
Smart advertisement dissemination: Advertisement publisher can modify advertising strategy by recognizing implicit Community structures.
Continuous Sentiment Analysis: System provides a opportunity to investigate market's sentiment by tracking of viewers' transition among multiple communities.
Community Generation Process
Community of Interest
Application

Figure 2. System Architecture of Stream-Oriented Community Generation.

focuses on the community-level viewing history, rather than a personal-level viewing history, in order to detect changes in a community's interests. The effect of a TV program can be assessed by analyzing the community-level sentiments using this tracking mechanism, and this information can be used to design TV program guides. This system reveals the transitions in the sentiments of communities, i.e., how many viewers change their opinion and how drastically they do this over the duration of a program.

The remainder of this paper is structured as follows. Section 2 presents motivating example of our community generation. Section 3 summarizes the related work briefly. Section 4 describes fundamental concept and system architecture. Finally, Section 5 concludes this paper.

II. Motivating Examples

Our community-oriented community generation system's unique feature is community-of-interest-based communication mechanism that allows viewers to publish their comments about a TV program. This mechanism provides a context to the fragmented messages on the Web, which is an asynchronous medium, through TV, which is a synchronous medium. Thus, in our ITV model, TV —the 1st screen— provides contexts to the interaction on the internet-enabled second screen —smartphone or tablet. This mechanism enables the community generation system to track the sentiments of viewers about a TV program. This feature realizes the following two applications.

1) Smart advertisement dissemination: The conventional advertising strategies can be made more efficient by targeting users in a community according to their specific and common interests rather than age group or sex. Our system recognizes community structures on the basis of users' interests rather than explicitly defined community characteristics such as the follower/followee ratio. As our implicit community structures are highly dynamic and flexible, advertisement publishers can select a target community structure community structure such that the advertisement is more relevant to the targeted consumers. From users' perspective, the community structure can help the users to filter out irrelevant advertisements.

2) Continuous Sentiment Analysis: This system provides an opportunity to investigate the market's sentiment by tracking of viewers' transition among multiple communities. Our system provides "live" feedback from SNS by capturing the structure of implicit communities and their sentiment continuously. Even if the member of community is changing, the system tracks such transition and generates an appropriate community.

III. Related Work

Since the invention of the Internet TV (ITV), there have been many attempts to integrate TV and new media. Tuomi [5] modeled the ITV as a substitution for a PC connected to Internet. Several methods have been developed for simultaneously displaying a TV stream and the related Web content by recommending Web resources as per the content of the TV stream [6][7]. As an advertisement recommendation method for Mobile TV, Adany et al.[8] has proposed a sequential solution procedure and several heuristic algorithms for uncertain personal advertisement allocation. TV2Web [9] is a seamless cross-media user interface that can be moved between TV screens and Web pages by linking units and displaying them smoothly using zooming metaphors. Many TV-on-Web methods have been proposed for displaying detailed information or service-related TV programs [9]-[11]. Another approach to integrate TV with the Web is the interactive approach, wherein the contents or the topic of a specific program are defined by [6] exchanging information or chatting [9] with other viewers in real time. Further, human-computer-interaction systems for TVs, PCs, and mobile devices have been proposed in [12].

IV. Stream-Oriented Community Generation

Figure 2 shows the system architecture of the community-oriented community generation system. This system extracts sentiments and keywords by analyzing a viewer's messages about a TV stream. The extracted keywords and sentiment information are integrated with the TV stream and its details obtained using an interactive program guide (IPG). This process provides “contexts” to the Web resources. This context information is used to analyze the implicit community structure that does not involves any explicit IDs or tags such as hashtags in Twitter and group functions supported by Facebook. This section describes the fundamental architecture and functions of the stream-oriented community generation.

4.1 Data Structure

Our system uses three types of data structures.

● Text Messages Exchanged on SNS: Text messages are contents posted by SNS users. The system analyzes these text messages. A text message is a tuple consisting of timestamp information, and the user’s identifier, and the text content. A set of messages is used as a primary data structure for sentiment analysis and consists of messages. Each message in a set of message is sequentially ordered according to the timestamp information in each message.

● TV Stream: The TV stream data is prefetched from an IPG and has information of the start and end time of a TV program respectively and keywords used for annotating the TV program. The annotation consists of the title, contents, performers, and word groups related to the TV program.

● Community: The system evaluates the word frequency and the sentiment information in each message in order to generate this community dynamically according to the messages or information posted by users sharing the same or similar interests. A community consists of the user ID derived from the SNS user set and keywords. Keywords include nouns and adjectives that represent users’ sentiments. The system deals with the community data and text massage data as vectors in a feature vector space. Our feature vector space consists of keywords describing the TV contents and SNS messages and is a typical high-dimensional vector space. The system employs a dynamic feature selection mechanism to create the appropriate context vector space for measuring the distance between (1) a community and a message, (2) two communities, and (3) two messages.

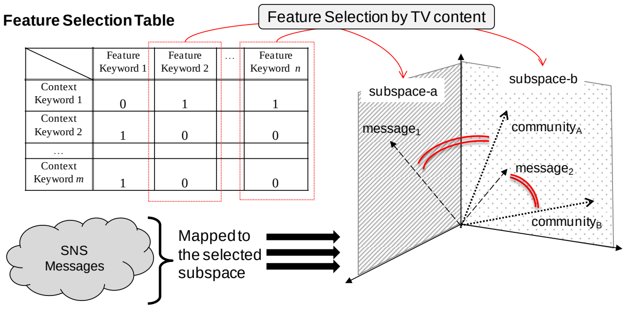
4.2 Dynamic Feature Selection for Analyzing Communities

In order to analyze the implicit communities, our system employs a dynamic feature selection mechanism that creates a low-dimensional vector space for each community. As a fundamental data structure, this system provides a high-dimensional vector space (e.g., 3,000 dimensions) that consisting of the necessary and sufficient number of feature keywords. This full-space can be used to represent various messages and communities, but it is difficult to precisely identify the data items in the space. Each dimension in the full-space corresponds to a specific keyword corresponding to the topics in TV stream. Our concept of feature selection is that a community should be generated by considering its own context, because a different context gives a different meaning in the social network. The system does not define the relevance between a context and a community statically.

Thus, this system creates sub-space according to the community. Figure 3 shows the results of a feature selection table that defines a “context keyword” as a trigger for selecting features. This table defines the correlation between the context keyword with topic made through TV program and the feature keyword derived from the messages exchanged through an SNS. This function eliminates the ambiguity of context that is caused by summarizing an entire users’ context into a feature vector.

TABLE I. FEATURE SELECTION TABLE

	Feature Keyword 1	Feature Keyword 2	…	Feature Keyword *n*
Context Keyword 1	0	1	⋯	1
Context Keyword 2	1	0	⋯	1
…				
Context Keyword *m*	0	0	⋯	0



	Feature Keyword 1	Feature Keyword 2	…	Feature Keyword *n*
Context Keyword 1	0	1		1
Context Keyword 2	1	0		0
…				
Context Keyword *m*	1	0		0

Figure 3. Feature Selection Operation for Creating Low-dimensional Vector Space to analyze Community Structure.

TABLE I shows the feature selection table. Here, the feature keyword is generated using a Web dictionary and has an enough capability to explain a user’s sentiment and interest. Context keywords play a role to define the context and to recall a set of feature keywords in a specific context. The feature selection process is as carried out follows:

● Step-1: Find the context keyword from the IPG data, and generate a weighted term-frequency context vector.

● Step-2: Transmit the context vector to the feature vector space by using the feature selection table. This transmission function $f_{map}(v,x)$ that inputs a context vector v and feature selection table x is defined as follows:

$$f_{map}(v,x) := \left(\sum_{i=0}^{m} v_{[i]} \cdot x_{[0,i]}, \cdots, \sum_{i=0}^{m} v_{[i]} \cdot x_{[n,i]} \right)$$

where $v_{[i]}$ is the value of the i -th context keyword and $x_{[n,i]}$ is the value of the n-th feature keyword corresponding to the i-th context keyword. n denotes the number of context keywords in the full-space. The context feature weights are converted into feature keyword weights by using this function.

ISBN: 978-1-61208-200-4

● Step-3: Select the top-k relevant feature keywords as the axis of the subspace, and compute the inner product between the vectors of community or SNS messages. This function $f_{distance}(r_1, r_2)$ that calculates the relevance between r_1 vector and r_2 vector is defined as follows:

$$f_{distance}(r_1, r_2) \coloneqq \sum_{i=0}^{n} r_{1[\mathrm{i}]} \cdot r_{2[\mathrm{i}]}$$

where r_1 and r_2 denotes a vector of community or a SNS message, and n denotes the number of features in the full-space.

● Step-4: Map every SNS message and community in the created subspace to compute the community-based relevance. The following function $f_{distance_{sub}}({r_1}', {r_2}')$ is used to calculate the relevance by using the sub-space:

$$f_{distance\_sub}({r_1}', {r_2}') \coloneqq \sum_{i=0}^{q} {r'}_{1[\mathrm{i}]} \cdot {r'}_{2[\mathrm{i}]}$$

where ${r_1}'$ and ${r_2}'$ denote vectors mapped into the subspace and q denotes the number of features in the subspace.

4.3 Function for Analyzing Implicit Communities

The main functions of this system can be divided into the following four categories: (1) time interval selection according to the context of the TV program, (2) community generation to create groups of users that share the same or similar interests, (3) community-of-interest(CoI)-based messaging, and (4) community transition analysis.

● Time Selection Function: Using the time interval selection function, the system extracts a set of messages from all the messages according to the time window defined for the TV program.

● Community generation Function: A set of communities is generated by morphologically analyzing the keywords for the current TV program. The community generation function generates communities according to the set of messages selected by the time section function and the candidate community to be mapped to the vector space.

● Community-of-Interest-based Messaging-Function: Using this function, the system delivers messages about a TV program to the relevant users. This messaging function involves a rather complex transfer process but provides users with an open and easy-to-use messaging platform. In concrete terms, the system selects target community from the set of all communities, and distributes the target community a SNS message.

V. CONCLUSION AND FUTURE WORKS

This paper proposes a stream-oriented community generation that associates real-time TV streams and Web resources by analyzing the communication among users on an SNS. Our approach realizes a novel communication medium that integrates messaging systems and TV streams according to viewers' interests. As a future work, we plan to implement this system on the 2nd screen (tablet devices) and perform experimental studies to demonstrate the effectiveness of our approach. We will also extend a range of association by not only focusing on the community but also handling other levels of communication scope, such as "Community & Individual Viewer," "TV-Stream & Web-Resources," "Individual Viewer & TV-Stream," and so on.

ACKNOWLEDGMENT

This research was supported by SCOPE: Strategic Information and Communications R&D Promotion Programme of Ministry of Internal Affairs and Communications, Japan: "Kansei Time-series Media Hub Mechanism for Impression Analysis/Visualization Delivery Intended for Video/Audio Media."

REFERENCES

[1] Ceasar, P. and Chorianopoulos, L. "Interactivity and user participation in the television lifecycle:creating, sharing and controlong content," Proc.UXTV 2008, pp. 127-128, 2008.

[2] Jensen, J., Tscheligi, M., Obrist, M., and Lugmayr, A. "Changing Television Environments," LNCS5066. Springer, pp. 1-10, 2008

[3] Ursu, M.F., Sussner, J., Myrestam, U., Hall, N., Thomas, M., Kegel, I., Williams, D., Tuomola, M., Lindstedt, I., Wright, T., Leurdijk, A., and Zsombori, V "Interactive TV narratives : Opportunities, progree and challenges," ACM TOMCCAP, Vol.4, Issue4, pp. 1-39, 2008.

[4] Donnelly, V. and Merrick, P. "Community portals through community generation," SIGCAPH Comput. Phys. Handicap, pp. 9-14, 2002.

[5] Tuomi, P. "A brief history of social iTV entertainment," In Proceedings of the MindTrek '09, ACM, pp. 15-18, 2009.

[6] Ma, O. and Tanaka, K. "Webtelop: Dynamic tv-content augmentation by using web pages," In Proceedings of ICME' 2003, IEEE, Vol. 2, pp. 173-176, 2003.

[7] Nadamoto, A. and Tanaka, K. "Complementing Your TV-Viewing by Web Content Automatically-Transformed into TV-program-type Content," In Proceedings of the ACM Multimedia' 2005, pp. 41-50, 2005.

[8] Adany, R., Kraus, S., and Ordóñez, F. "Uncertain personal advertisement allocation for Mobile TV," In Proceedings of ACM MoMM '10, pp. 159-166, 2010.

[9] Sumiya, K., Munisamy, M., and Tanaka, K. "Tv2Web: generating and browsing web with multiple lod from video streams and their metadata," In Proceedings of ICKS2004, pp. 158-167, 2004.

[10] Miyamori, H. and Tanaka, K. "Webified Video: Media Conversion from TV Programs to Web Content for Cross-Media Information Integration," In DEXA2005, LNCS3588, Springer, pp. 176-185, 2005.

[11] Livingston, K., Dredze, M., Hammond, K., and Birnbaum, L. "Beyond Broadcast," In Proceedings of ACM IUI'2003, The Seventh International Conference on Intelligent User Interfaces, pp. 260-262, 2003.

[12] Martin, R. and Holtzman, H. "Newstream: A Multi-Device, Cross-Medium," and Socially Aware Approach to News Content. In Proccedings of the ACM EuroiTV 2010, pp. 83-90, 2004.

Query-by-Appearance System for Style-Oriented Media Retrieval

Yuka Koike
Faculty of Environment and Information Studies
Keio University
5322 Endo, Fujisawa, Kanagawa, 252-0882, Japan
t09334yk@sfc.keio.ac.jp

Shuichi Kurabayashi
Faculty of Environment and Information Studies
Keio University
5322 Endo, Fujisawa, Kanagawa, 252-0882, Japan
kurabaya@sfc.keio.ac.jp

***Abstract*— This paper proposes the *Query-by-Appearance* system that provides an intuitive and effective query-input and visual retrieval method for media data, especially e-books, based on similarity of content's layout and color. The unique feature of this system is a query-assistance function that helps users to input their requirements by utilizing the knowledge which have been developed in the field of "Editorial Design". Editorial design is an essential methodology to enrich overall appearance of books and magazines. The query input assistant retrieves editorial design templates that are similar to the input query, and generates images by combining these templates with the chosen color scheme. The system then retrieves actual e-book images that are similar to those generated from the query. Finally, the system visualizes the retrieval results, which consist of two relevance scores (layout and color) in a two-dimensional ranking plot. This assistant mechanism allows users to find the desired e-book by submitting a query, which consists of simple lines, intuitively.**

***Keywords*— e-Book; Search Engine; Editorial Design.**

I. INTRODUCTION

The proliferation of portable, personal devices dominated by a large display, such as tablet computers and smartphones [1][2] has made it common for e-books and Web pages to viewed on such devices. E-books, in particular, are fast spreading; Association of American Publishers reported that the total e-book sale from January to October of 2010 constituted 8.7% of all book sales in the United States [3]. Such proliferation and diversity of digital media data increase the demand of a system for retrieving them. Generic retrieval methods, such as keyword-based search engines and content-based image retrieval systems, are not sufficient for retrieving them, because users require highly-domain-specific search quality in such daily-life media data. For example, a fashion magazine requires a fashion-domain specific query for retrieving it, and an outdoor and nature magazine requires different query for retrieving them. It is essential to provide a query which is tailored to each type of e-books.

However, common users have difficulty in learning a new and specific search method for each type and genre of media data. It is important to develop a novel and intuitive search engine for those daily-life multimedia data. Visual-oriented search mechanisms, which do not use text-based search methods, are promising because users often memorize e-book contents by associating them with their visual appearances. We suppose that the overall visual appearance of page layout is essential to e-book search. In spite of layout being an important factor for bookbinding, there are no studies about layout searching method for e-book, because conventional systems utilize CBIR (Content-Based Image Retrieval) mechanisms for retrieving e-book.

Toward the above objective, this paper proposes a "*Query-by-Appearance*" system that improves the query input process by exploiting the knowledge of book design. This system provides a query-generation mechanism that enhances and decorates a user's rough and simple query by generating candidates of queries inspired by the initial input. This system enables users to find a desired e-book just inputting intuitive and simple query alone because the system ranks e-books according to the similarity of overall compositions, layouts, colors, and overviews, rather than detailed and trivial differences between them. We call this visually rich search method "*Query-by-Appearance*".

The unique feature of this system is a book style-oriented query generation, which is enhanced by the knowledge of "editorial design", for helping users to input a complex and sophisticated query by generating candidates of queries. Editorial design, developed in the book industry, is an essential methodology to enrich and to beautify books and magazines by configuring the overall appearance, including the visual layout, such as photographs and illustrations. Editorial design techniques provide sophisticated way to construct and to lay out visual objects' compositions. Our "*Query-by-Appearance*" system utilizes the editorial design methods as query templates. The query templates consist of layouts and color schemes. When the system receives a query, which consists of simple lines, the system applies the query templates to enrich and to improve the query, and generates the well-organized and colored image query by combining the layout templates and the color scheme templates.

We design "*Query-by-Appearance*" system as an embedded system in an e-book format, such as EPUB3. The system encapsulates the templates, which are derived from the editorial design methods, into each e-book according to the style of books such as a fashion magazine and a mystery novel. This design principle makes it possible to interpret a user's query with reflecting the target books' own characteristics. A core function of this system is a design-based template-matching function that compares a user's rough query input with the embedded editorial design templates.

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-200-4

The system is applicable to the following area: 1) searching dimly remembered visual contents, such as the title of book, web pages, and 2) searching jackets of DVD, CD, and packaging of a product by preparing appropriate templates. We show a prototype system implementation that is embeddable into a media data, by using JavaScript supported in EPUB3 and HTML5. This prototype system implements self-search mechanism in those media data according to the internal page images and images of web pages in the browser history.

The remainder of the paper is structured by follows. In section two, we discuss several related studies. In section three, we show an architectural overview of our system. In sections four and five, we describe the four fundamental components and three core functions of our system. In section six, we show the prototype of our system. In section seven, we evaluate the effectiveness of our system. In section eight, we give concluding remarks.

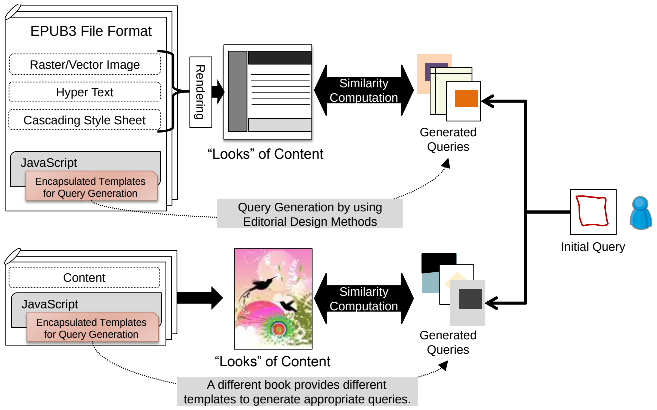


Figure 1. A Concept of Query-by-Appearance System for Style-Oriented e-Book Retrieval by Using Encapsulated Editorial Design Templates for Query Generation.

II. Related Works

There are several studies on developing a search engine that considers the visual appearance of media. Recently, CBIR [4][5] systems have been developed that allow users to search for images by visual similarity. Such systems support "*Query-by-Example*," which requires an example image as the query input. However, basically, CBIR methods are not suitable to find complex multimedia data such as Web sites and e-books because these methods ignore the semantic information associated with the images, such as text and annotations. Alternatively, a method [6] has been proposed to calculate similarity among the visual components of Web pages by their overall visual appearance. The system proposed in [7] analyzes the visual link structure created by assigning numerical weights to each image. This system incorporates visual signals into text-based search engines in order to improve the accuracy of conventional search engine. Those conventional approaches are effective to retrieve image data by submitting the detailed query, such as a sketch, an example image, and a combination of them. However, it is difficult for novices to input such as detailed query, especially in the domain of well-designed e-books. Thus, a new approach to assist users to create the domain-specific query for multimedia data is required.

III. System Architecture

Figure 2 illustrates an architectural overview of this system. Our "*Query-by-Appearance*" system provides an indirect e-book retrieval mechanism that is leveraged by the knowledge base of editorial design. This system retrieves the e-book images that are similar to editorial design characteristics inspired by the query. The results are visualized in two-dimensional ranking, such as generated query axis and ranking axis. This query-assistance function enables users to input layout sketch and color by using Web-based touch UI, and search e-books through rough visual appearances intuitively, by accepting a simple query consisting of layouts and colors, rather than technical knowledge.

The system uses the generated e-book image data for retrieving an e-book by calculating similarity based on the color and structure of appearance. This approach is highly effective to develop a Web-based touch UI because this system simplifies the query-input task. The touch UI visualizes both the layout similarity and the color similarity in a two-dimensional matrix. This visualization enables users to retrieve the desired e-book according to their preferences, which means weights of the layout similarity and the color similarity, by selecting sub-matrix in the visualized two-dimensional matrix.

The most important function of this system is the query-assistance function that utilizes both layout and color to enrich and extend the input query. As described, this system employs a design template database which stores layouts and color scheme data. The templates are defined by sets of matrices; so, the system is able to select the relevant templates by calculating the similarity in layout structure. In addition, 20 color schemes are provided by the combination of 102 colors.

This system performs the *Query-by-Appearance* for e-books by involving the following six steps:

1. The user inputs a rough layout sketch on the Canvas system.
2. Second, the system converts the rough layout query input into a matrix, and compares this matrix with 30 templates that are stored in the knowledgebase.
3. The system generates colored templates by adding color to the selected layout templates using the 20 color schemes stored in the knowledgebase. The sets of colored templates are expanded automatically to sets of image matrices, which include the layout structure and the color scheme.
4. The system calculates similarity between the expanded query matrices and e-book image data stored in an e-book database. The e-book image data is preliminarily clustered into 102 colors.
5. The e-book cover images that contain similar editorial design characteristics are visualized in a two-dimensional ranking with generated query axis and ranking axis.

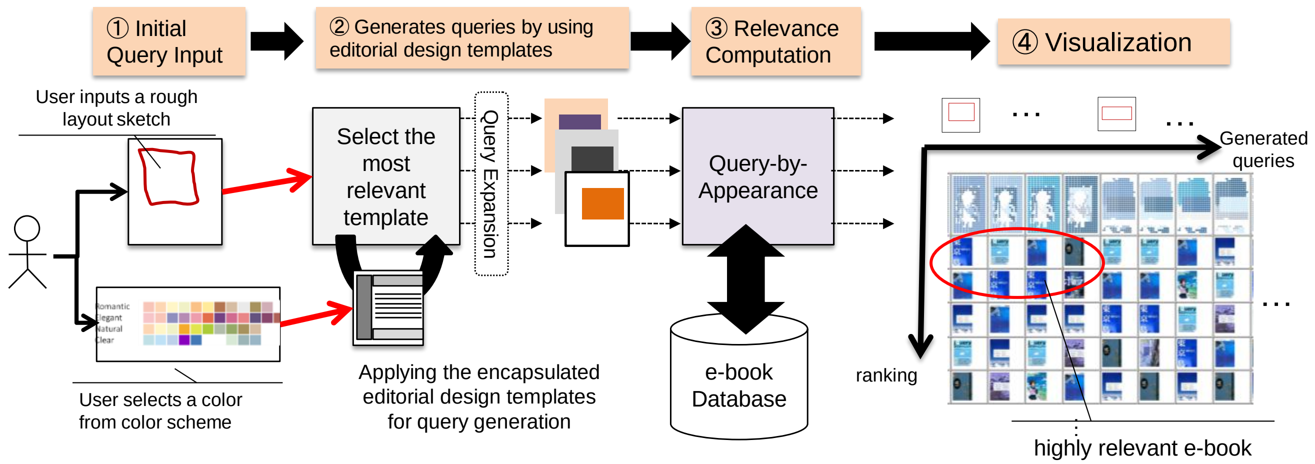

① Initial Query Input
② Generates queries by using editorial design templates
③ Relevance Computation
④ Visualization
User inputs a rough layout sketch
Select the most relevant template
Query Expansion
Query-by-Appearance
Generated queries
Romantic
Elegant
Natural
Clear
e-book Database
ranking
User selects a color from color scheme
Applying the encapsulated editorial design templates for query generation
highly relevant e-book

Figure 2. System Architecture of Query-by-Appearance System.

6. Further retrieval is done by selecting a color on a sub-matrix in the visualized two-dimensional matrix.

Our visualization mechanisms allow users to retrieve the desired e-book according to similarity in layout and color scheme.

IV. Data Structures

Our system contains four fundamental components: A) Query Matrix, B) Template Knowledgebase, C) e-book Database, and D) Visualization.

A. *Query Matrix*

Our system converts each query into a matrix that represents the user`s rough input, such as simple lines drawn on HTML5 Canvas. We call this matrix a query matrix. The query matrix is important for calculating the similarity between the user input and the layout templates in the knowledgebase. We assign weights to each column as follows: if a line passes through a column, 1.0 is assigned to the column; if no line passes through a column, 0.0 is assigned to it. This weight is assigned to get the inner product of the user`s query matrix and layout template matrix, as shown in the *Template Knowledgebase*.

The query matrix Q is defined as follows:

$$\mathrm{Q} \coloneqq \begin{bmatrix} w_{[0,0]} & \cdots & w_{[0,m]} \\ \vdots & \ddots & \vdots \\ w_{[n,0]} & \cdots & w_{[n,m]} \end{bmatrix}$$

where $w_{[n,m]}$ indicates whether a line is present at $[n,m]$ in the query.

B. *Template Knowledgebase*

The system provides matrix structures representing: 1) the layout template 2) the assigned-color template, and 3) the color scheme. We call the data structures a design template structure. All of the above are stored in the Template Knowledgebase, which assists in expanding the input queries.

1) The layout template matrix is the set of matrices obtained by exploiting the method of editorial design. We assign weights to each column as follows: if there is most obvious layout border passes through a column, 1.0 is assigned to that column; the regions surrounding 1.0, 0.5 is assigned to it; if there is no important layout structure to the column, 0.0 is assigned to it.

The layout template matrix $\mathrm{T_L}$ is defined as follows:

$$\mathrm{T_L} \coloneqq \begin{bmatrix} t_{[0,0]} & \cdots & t_{[0,m]} \\ \vdots & \ddots & \vdots \\ t_{[n,0]} & \cdots & t_{[n,m]} \end{bmatrix}$$

where $t_{[n,m]}$ indicates whether a line is present at the location $[n,m]$ in the template matrix.

2) The assigned-color template matrix is a set of matrices containing numbers from 1 through 5 in each column. The system draws color in each column according to this number. 1 represents the color that appears most often in the image, whereas 5 represent the color that appears least often.

3) The color scheme is a set of color data obtained by a combination of 102 colors. This system draws colors in each column by using this color scheme, to perform the expanded queries in order to generate e-book image data.

The color scheme $\mathrm{T_C}$ is defined as follows:

$$\mathrm{T_C} \coloneqq \langle \mathrm{c_1}, \mathrm{c_2}, \cdots, \mathrm{c_k} \rangle$$

where $\mathrm{c_k}$ denotes each color used in the color scheme, and k is the total number of colors in the color scheme. In the examples in Figure 3, each color scheme is assigned a specific adjective.

20 color schemes are made using a dictionary that defines adjectives for color schemes, such as “Romantic”, or “Elegant” [8].

C. *e-book Database*

Our system converts JPEG-format image data into the matrix structure by clustering the image into 102 colors. We call the data structure an e-book database matrix. This e-book database matrix contains clustered image data that is used to calculate the similarity with the expanded query matrix obtained from the user input and template knowledgebase.

The e-book database matrix D is defined as follows:

$$\mathrm{D} := \begin{bmatrix} d_{[0,0]} & \cdots & d_{[0,m]} \\ \vdots & \ddots & \vdots \\ d_{[n,0]} & \cdots & d_{[n,m]} \end{bmatrix}$$

where d represents each element of the matrix, and [n, m] represent the rows and columns of the image data, respectively, in the matrix.

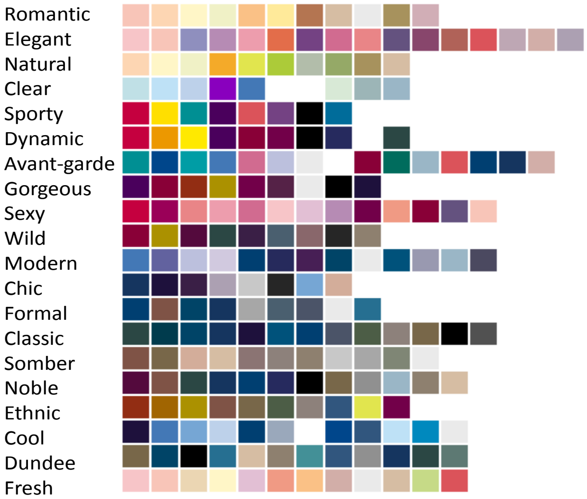


Figure 3. Sets of color schemes.

D. Visualization

The system visualizes the layout and color similarity in a two-dimensional ranking. This ranking visualization consists of two relevance scores for each cell. The ranking array $\mathrm{Cell_x}$ is defined as follows:

$$\mathrm{Cell_x} := \langle \mathrm{L_{i,}C_{j,}R_{L,}R_{C,}}\{\mathrm{B_1}\cdots\mathrm{B_p}\}\rangle$$
$$\vdots$$
$$\mathrm{Cell_{x+n}} := \langle \mathrm{L_{i,}C_{j,}R_{L,}R_{C,}}\{\mathrm{B_1}\cdots\mathrm{B_p}\}\rangle$$

where $\mathrm{Cell_x}$ to $\mathrm{Cell_{x+n}}$ comprise $\mathrm{L_i}$ for Layout id, $\mathrm{C_j}$ for color id, $\mathrm{R_l}$ for layout ranking, $\mathrm{R_c}$ for color ranking, and $\{\mathrm{B_1}\cdots\mathrm{B_p}\}$ for ranking order on a column, when more than one image data has the same layout template and color scheme. The order of books in a certain column is decided by color relevancy, where p represents the order of the books.

V. CORE FUNCTIONS

The system contains three fundamental functions: A) Selecting the most relevant template, B) Assigning-colors to the template C) Query-by-Appearance, and D) Two-dimensional ranking.

A. Selecting the most relevant template

The system selects the most relevant template by comparing the user input with the template matrices stored in the template knowledgebase. We call this function query assistance. The inner product of the transformed user input matrix, which consists of line sketched data, and the template matrices that are stored in the template knowledgebase, is calculated in order to analyze the similarities. A higher value for this inner product indicates greater similarity.

The function is defined as follows:

$$f_{layout}(q,t) := \sum_{i=1}^{w}\sum_{j=1}^{h} q_{[i,j]}t_{[i,j]}$$

where $q_{[i,j]}$ indicates whether a line is present at position [i,j] in the query matrix, and $t_{[i,j]}$ indicates whether a line is present at the position [i,j] in the template matrix. This calculation measures the line similarity between the query matrix and the template matrix by checking all the pixels.

B. Assigning-colors to the template

The system assigns colors 1 through 5 to the template, according to the users selected color scheme and number of colors. This function is important for generating a colored template, which contains template, keyword, and number information. The function is defined as follows:

$$f_{assign}(T_L, \mathrm{keyword}, \mathrm{number}) \rightarrow \mathrm{Q} := \begin{bmatrix} hsv_{[0,0]} & \cdots & hsv_{[0,m]} \\ \vdots & \ddots & \vdots \\ hsv_{[n,0]} & \cdots & hsv_{[n,m]} \end{bmatrix}$$

where template $\mathrm{T_L}$ indicates the template data that contains numbers 1 to 5 for drawing, a keyword indicates the name of the color scheme, and number indicates the number of colors in color scheme to draw on the template.

C. Query-by-Appearance

Our system calculates the correlation between the generated query matrix and e-book image data matrices in order to retrieve the relevant e-book image. The Query-by-Appearance operation is defined as follows:

$$f_{qba}(Q,B) := \sum_{i=1}^{w}\sum_{j=1}^{h} \Delta_{godlove}\left(Q_{[i,j]}, B_{[i,j]}\right),$$

$$\Delta_{godlove} := \frac{2S_1S_2\left(1-\cos\left(2\pi\frac{|H_1-H_2|}{100}\right)\right)+(|S_1-S_2|)^2+(4|V_1-V_2|)^2}{2}$$

where $Q_{[i,j]}$ denotes a color at the specific point in the generated query matrix, and consists of HSV components $\mathrm{H_1}$, $\mathrm{S_1}$, and $\mathrm{V_1}$. $B_{[i,j]}$ denotes a color at a specific point in the e-book image data, and consists of $\mathrm{H_2}$, $\mathrm{S_2}$, and $\mathrm{V_2}$. This calculation measures the color similarity between the expanded queries and the e-book image data for each block. We employ Godlove's delta equation [9] to calculate the distance between two HSV colors.

D. Two-dimensional ranking

Our system provides a two-dimensional visualization mechanism for presenting the results of the e-book image search. This visualization enables users to select desired books on the basis of the layout and the color similarity. The system takes a wider sample of template data compared to the user`s input, in order to display the retrieval results interactively on the basis of the choice of layout and color.

This system employs the following three steps:

Step-1. The generated queries are shown on the horizontal axis. Each column contains a query with template and color.

Step-2. The ranking of e-book image data is shown on the vertical axis.

Step-3. The system stacks the e-book image data according to the similarity score of queries and e-book meta data in each cell.

VI. SYSTEM IMPLEMENTATION

We have implemented a prototype system for *Query-by-Appearance* that calculates the similarity between generated queries and editorial design templates. The prototype system is implemented using HTML5 Canvas and JavaScript as shown in Figure 4. The important feature of this implementation is that it uses standard web technologies, which are also used by the current EPUB3 specification. So, we can apply our method to EPUB3-based e-books without significant modification. The implemented system consists of the following three modules:

- View: The user sketches the layout query on HTML5 Canvas and selects a color scheme. The system calculates the similarity between the queries and template knowledgebase.
- Controller: The retrieval is done by clicking on the search button.
- Model: The retrieval target database and two templates (layout matrix, color scheme) are encoded in the JSON format and embedded in the HTML5.

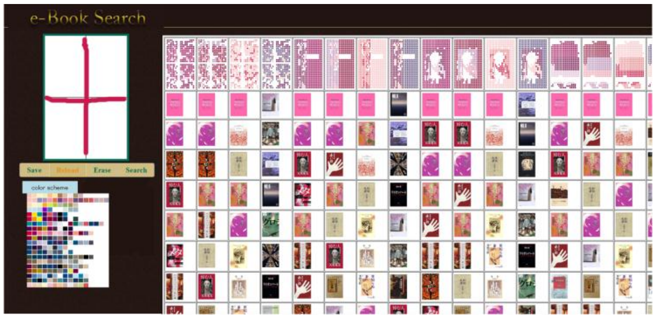


Figure 4 A Visualization of e-book search engine.

VII. EVALUATION

A. *Outline of experimental studies*

In this section, we evaluate the effectiveness of our system by examining the retrieval precision for our generated queries. The task of this experiment is to clarify the effectiveness of retrieving book cover image by utilizing the knowledge of “editorial design”. We compare the retrieval precision using two search methods as follows: method-1) uses queries generated using only layout templates, and method-2) uses queries generated by integrating layout templates and color scheme templates. We show that the integration of a layout template and a color template has significant contribution to the e-book retrieval result.

In this experiment, we have prepared: 1) 300 book cover images from Amazon.co.jp, 2) five queries, 3) five answer data sets for each query. The queries and answer data sets are set up considering the basic structure of “editorial design”, as specified as follows: 1) the symmetrical layout, 2) the diagonal layout, 3) the layout with gravity point on center. We chose the following five queries (Figure 5):

Query-1. Draw the shape of a cross to divide the canvas into four sections.

Query-2. Draw a vertical line along the center of canvas to divide it into two sections.

Query-3. Draw two horizontal lines on the canvas to divide it into three sections.

Query-4. Draw a rectangle in the center of the canvas to divide it into two sections.

Query-5. Draw two vertical lines on the canvas to divide it into three sections.

We have used color schemes that consist of four colors (as shown in Figure 3). The system calculates the similarity between those queries and generated e-book image data.

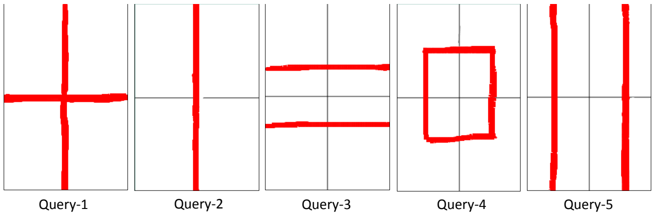


Figure 5. The Experimental Queries drawn on the Canvas.

B. *Experimental Results*

In this section, we have evaluated the retrieval precision of generated queries in order to clarify the effectiveness of our approach. Figure 6 shows the results of the retrieval. Our approach gives two average scores for retrieved rank. The first score is the average of correct answer data shown in the top 20, which has retrieved only by template relevance. The second score is the average rank of correct answer data that has the chosen color scheme.

Figure 7 shows the resulting scores of this experiment. The vertical axis is the retrieved score. The horizontal axis is the variations of the query that is the template of correct answer sets. We calculate the average scores of each query in the top 20 ranking as shown in Figure 6. This result shows that method-2 (integrating both layout and color templates) achieves better retrieval precision than method-1 (using the layout template only), since a superior rank has been assigned to the result corresponding to the query. This result shows that the system effectively retrieves e-book images by using layout template and color scheme.

VIII. CONCLUSION AND FUTURE WORKS

This paper proposes a “*Query-by-Appearance*” system for e-books, which provides an intuitive and effective query input and visual retrieval method base on the similarity in overall layout and color scheme. The unique feature of this system is the query-assistance function that exploits structural design and analysis knowledge from the field of “editorial design”. Editorial design is an essential methodology to enrich the appearance of books and magazines. The query input assistant retrieves the editorial design templates that are similar to the input query, and generates actual e-book images by combining the color templates and the layout templates. Then, the system retrieves e-book images that are similar to the generated e-book image. Finally, the system visualizes the retrieval results, which consist of two relevance scores (layout, color) in a two-dimensional ranking; generated query axis and ranking axis. This assistant mechanism is intuitive because it allows users to find a desired e-book by submitting a query that consists of simple lines.

As a future work we are planning to develop a prototype system that supports full-spec EPUB3, and to perform a feasibility study by evaluating scalability and effectiveness of our approach applied to the existing e-books.

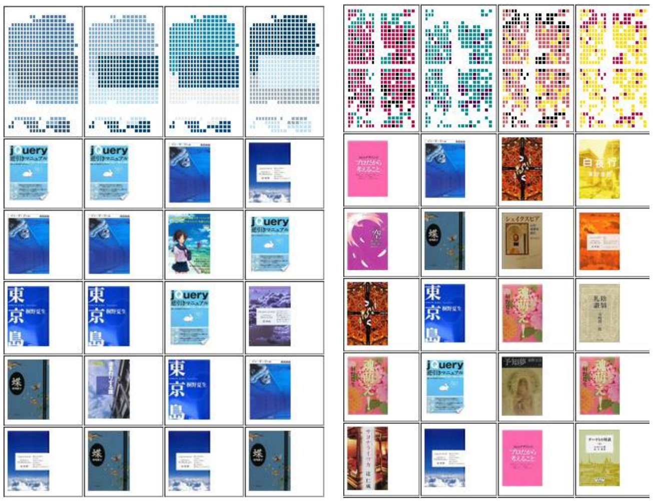


Figure 6. The results of top 20 retrieved images. The combination of Query 3 and "cool"(left), and Query 1 and "sporty"(right).

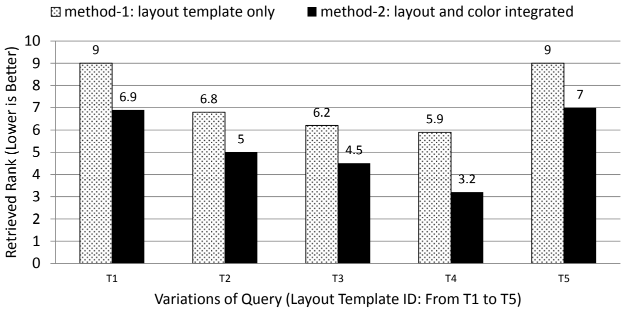


Figure 7. The result of retrieved rank of "only layout template" and "colored template".

ACKNOWLEDGMENT

This research is supported by SCOPE: Strategic Information and Communications R&D Promotion Programme of the Ministry of Internal Affairs and Communications, Japan: "Kansei Time-series Media Hub Mechanism for Impression Analysis / Visualization / Delivery Intended for Video / Audio Media".

REFERENCES

[1] Gartner, "Gartner Says Worldwide PC Shipments in First Quarter of 2011 Suffer First Year-Over-Year Decline in Six Quarters", 13 April 2011, http://www.gartner.com/it/page.jsp?id=1632414.

[2] IDC, "Nearly 18 Million Media Tablets Shipped in 2010 with Apple Capturing 83% Share; eReader Shipments Quadrupled to More Than 12 Million, According to IDC", 10 March 2011, http://www.idc.com/about/viewpressrelease.jsp?containerId=prUS22737611.

[3] American Association of Publishers, "AAP Reports October Book Sales", 2010, http://www.publishers.org/main/PressCenter/Archicves/2010_Dec/AAPReportsOctoberBookSales.htm, (accessed 2010-9-10).

[4] Baeza-Yates, R. and Ribeiro-Neto, B. "Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)," Chapter 14, Addison-Wesley Professional, 2011.

[5] Datta, R., Li, J., and Wang, J. Z. "Content-Based Image Retrieval - Approaches and Trends of the New Age," in Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval, ACM: New York, NY, USA, 2005.

[6] Furusawa, T., Watai, Y., Yamasaki, T., Aizawa, K. "A Visual Similarity Metric for View-based Web Page Retrieval," The Journal of the Institute of Image Information and Television Engineers, Vol. 62, No.2, 2008, pp.209-215, DOI: 10.3169/itej.62.209 (In Japanese).

[7] Jing, Y., and Baluja, S. "PageRank for Product Image Search," In Proceedings of the 17th International Conference on World Wide Web (2008), pp. 307-316, April 2008.

[8] Nagumo, H. "Color Scheme Imaging", ISBN-13: 978-4766111705, Graphics-sha, 2000 (In Japanese).

[9] Godlove, I. H. "Improved Color-Difference Formula, with Applications to the Perceptibility and Acceptability of Fadings". J. Opt. Soc. Am., Vol. 41, NO. 11, pp.760-770, 1951.