ICIW 2013

The Eighth International Conference on Internet and Web Applications and Services

ISBN: 978-1-61208-280-6

June 23 - 28, 2013

Rome, Italy

ICIW 2013 Editors

Ioannis Moscholios, University of Peloponnese - Tripolis, Greece

Marek Rychly, Brno University of Technology, Czech Republic

ICIW 2013

Forward

The Eighth International Conference on Internet and Web Applications and Services (ICIW 2013) held on June 23 - 28, 2013 - Rome, Italy, 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 2013 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 2013 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 2013. 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 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICIW 2013 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. We also hope the attendees enjoyed the historic charm Rome, Italy.

ICIW 2013 Chairs

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

ICIW 2013

Committee

ICIW Advisory Committee

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

ICIW 2013 Technical Program Committee

Charlie Abela, University of Malta, Malta
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
Feda AlShahwan, The Public Authority for Applied Education and Training (PAAET), Kuwait
Eckhard Ammann, Reutlingen University, Germany
Liliana Ardissono, Universita' di Torino, Italy
Ezendu Ariwa, London Metropolitan University, UK
Khedija Arour, University of Carthage - Tunis & El Manar University, Tunisia
Johnnes Arreymbi, University of East London, UK
Marzieh Asgarnezhad, Islamic Azad University of Kashan, Iran
Jocelyn Aubert, Public Research Centre Henri Tudor, Luxembourg
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
Tharrenos Bratitsis, University of Western Macedonia, Greece
Maricela Bravo, Autonomous Metropolitan University, Mexico
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
Ana Regina Cavalcanti Rocha, Federal University of Rio de Janeiro, Brazil
Ajay Chakravarthy, University of Southampton, UK
Xi Chen, Nanjing University, China
Dickson Chiu, Dickson Computer Systems, Hong Kong
Gianpiero Costantino, Institute of Informatics and Telematics - National Research Council (IIT-CNR) of Pisa, Italy
María Consuelo Franky, Pontificia Universidad Javeriana - Bogotá, Columbia
Javier Cubo, University of Malaga, Spain
Roberta Cuel, University of Trento, Italy

Florian Skopik, AIT Austrian Institute of Technology, Austria
Vladimir Stancev, SRH University Berlin, Germany
Michael Stencl, Mendel University in Brno, Czech Republic
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, Isfahan University of Technology, Iran
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
Nikos Tsirakis, University of Patras, Greece
Pavel Turcinek, Mendel University in Brno, Czech Republic
Samyr Vale, Federal University of Maranhão - UFMA - Brazil
Dirk van der Linden, Artesis University College of Antwerp, Belgium
Perla Velasco-Elizondo, Autonomous University of Zacatecas, Mexico
Iván P. Vélez-Ramírez, Phidelix Technologies, Puerto Rico
Maurizio Vincini, Universita' di Modena e Reggio Emilia, Italy
Michael von Riegen, University of Hamburg, Germany
Alexander Wöhrer, Vienna Science and Technology Fund, Austria
Michal Wozniak, Wroclaw University of Technology, Poland
Rusen Yamacli, Anadolu University, Turkey
Zhixian Yan, Samsung Research America, USA
Sami Yangui, Telecom SudParis, France
Beytullah Yildiz, Tobb Economics and Technology University, Turkey
Amelia Zafra, University of Cordoba, Spain
Martin Zimmermann, Hochschule Offenburg - Gengenbach, Germany
Christian Zirpins, Karlsruhe Institute of Technology, Germany
Jan Zizka, Mendel University in Brno, Czech Republic

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Agile Model-Driven Modernization to the Service Cloud 1
Iva Krasteva, Stavros Stavru, and Sylvia Ilieva

Ubiquitous System to Enhance the Supply Chain 10
Candido Caballero-Gil, Jezabel Molina-Gil, Pino Caballero-Gil, and Alexis Quesada-Arencibia

Modding and Cloud Gaming: Business Considerations and Technical Aspects 14
Alexander Wohrer, Yuriy Kaniovskyi, and Maximilian Kobler

Three-Tiered Data Mining for Big Data Patterns of Wireless Sensor Networks in Medical and Healthcare Domains 18
Jong Yoon

A Novel Risk-based Approach for Online Community Management 25
Bassem Nasser, Vegard Engen, Simon Crowle, and Paul Walland

Enforcing Data Availability in Structured Peer-to-Peer Storage Systems With Zero Replica Migration 31
Mesaac Makpangou

Robust and Semi-automatic Electronic Health Record Dissemination Using the Devices Profile for Web Services 38
David Gregorczyk, Timm Busshaus, and Stefan Fischer

Semantic Web Services Adaptation and Composition Method 45
Hamid Mcheick and Amel Hannech

Towards a New Trust Model for Health Social Networks 52
Sojendra Pradhan and Valerie Gay

A Web Service Migration Framework 58
M. Mohanned Kazzaz and Marek Rychly

Web Service and Structure of University Data 63
Masaaki Ida

A Visual Semantic Search Framework for Finding Craft Services 69
Maximilien Kintz and Andrea Horch

A Requirements Model for Composite and Distributed Web Mashups 75
Vincent Tietz, Oliver Mross, Andreas Rumpel, Carsten Radeck, and Klaus Meissner

An Approach to Dynamic Discovery of Context-Sensitive Web Services 83

Victor Gomes da Silva, Carlos Eduardo Cirilo, Antonio Francisco do Prado, Wanderley Lopes de Souza, and Vinicius Pereira

A New Unsupervised Web Services Classification based on Conceptual Graphs 90
Eiman Boujarwah, Hamdi Yahyaoui, and Mohammed Almulla

Enhancing Semantic Web Services Discovery Using Similarity of Contextual Profile 95
Zahira Chouiref, Abdelkader Belkhir, and Allel Hadjali

Internet of Threads 100
Renzo Davoli

A Complexity Analysis of an XML Update Framework 106
Mohammed Al-Badawi and Abdallah Al-Hamadani

Portable Full-text Retrieval System 114
Takehiko Murakawa and Tatsuya Takehara

Enabling End Users to Build Situational Collaborative Mashups at Runtime 120
Gregor Blichmann, Carsten Radeck, and Klaus Meissner

Issues in Conducting Expert Validation and Review and User Evaluation of the Technology Enhanced Interaction Framework and Method 124
Kewalin Angkananon, Mike Wald, and Lester Gilbert

Chord-Cube: Multiple Aspects Visualization & Navigation System for Music by Detecting Changes of Emotional Content 129
Tatsuki Imai and Shuichi Kurabayashi

Compressing Large Size Files on the Web in MapReduce 135
Sergio De Agostino

A Semantically Enriched E-Tendering Mechanism 141
Jingzhi Guo and Ka-Ieong Chan

Social Query: A Query Routing System for Twitter 147
Cleyton Souza, Jonathas Magalhaes, Evandro Costa, and Joseana Fechine

Using Chaos Theory to Investigate the Development Process of the Most Popular Blog 154
Kwoting Fang and Shu Chuan Wang

Stigmergy within Web Modelling Languages: Positive Feedback Mechanisms 160
Aiden Dipple, Kerry Raymond, and Michael Docherty

Incorporating Flow Theory to Technology Acceptance Model for Online Community Formation 167
Mayank Sharma, Pradeep Kumar, Bharat Bhasker, and Abhijit Bhatacharya

Creativity Detection in Texts 174
Costin-Gabriel Chiru

StreamQuilt: A Timeline-Aware Integration of Heterogeneous Web Streams 181
Riho Nakano and Shuichi Kurabayashi

Designing Formulas for Creating a Healthcare Cloud Service Based on a Formula Calculation Platform Service 187
Xing Chen, Keiichi Shiohara, and Hikaru Tazumi

RESTful Correlation and Consolidation of Distributed Logging Data in Cloud Environments 194
Christian Pape, Sven Reissmann, and Sebastian Rieger

Synergic Data Extraction and Crawling for Large Web Sites 200
Celine Badr, Paolo Merialdo, and Valter Crescenzi

Rethinking Traditional Web Interaction 206
Vincent Balat

GuidedBased Usability Evaluation on Mobile Websites 212
Bahadir Dundar, Nejat Yumusak, and Samet Arsoy

A New Technology to Adapt the Navigation 218
Rim Zghal, Corinne Amel Zayani, and Ikram Amous

Optimal Malicious Agreement in a Virtual Subnet-based Cloud Computing Environment 224
Kuo-Qin Yan, Hsueh-Hsun Huang, Shu-Ching Wang, and Shun-Sheng Wang

The Anatomy Study of Load Balancing in Cloud Computing Environment 230
Shu-Ching Wang, Ching-Wei Chen, Kuo-Qin Yan, and Shun-Sheng Wang

Cross-Media Retrieval for Music by Analyzing Changes of Mood with Delta Function for Detecting Impressive Behaviours 236
Yoshiyuki Kato and Shuichi Kurabayashi

Synote Discussion: Extending Synote to Support Threaded Discussions Synchronised with Recorded Videos 240
Mike Wald, Yunjia Li, Ea Draffan, James Brierley, Alyona Ivanova, Robert Streeting, and Matthew Tucker

A Model-Driven Approach for Service Oriented Web 2.0 Mashup Development 246
Jose Luis Herrero Agustin, Pablo Carmona, and Fabiola Lucio

Digital Badges in Informal Learning Environments 252
Bradley Barker

A Mobile Learning App for Driving Lessons 256
Jezabel Molina-Gil, Candido Caballero-Gil, Pino Caballero-Gil, and Alexis Quesada-Arencibia

Monitoring Activities in an E-Learning 2.0 Environment: A Multi-Agents system 260
Henda Belaid-Ajroud, Benedicte Talon, and Insaf Tnazefti-Kerkeni

A Multi-agent System to Implement a Collaborative Learning Method 266
Hanaa Mazyad, Insaf Tnazefti-Kerkeni, and Henri Basson

Agile Model-Driven Modernization to the Service Cloud

Iva Krasteva
Rila Solutions EAD
Acad. G. Bonchev str., bl. 27
Sofia, Bulgaria
ivak@rila.bg

Stavros Stavru
Faculty of Mathematics and Informatics, Sofia University
5, James Boucher Blvd
Sofia, Bulgaria
stavross@fmi.uni-sofia.bg

Sylvia Ilieva
IICT-BAS
Acad. G. Bonchev str., bl. 25A
Sofia, Bulgaria
Sylvia@acad.bg

***Abstract*— Migration of legacy systems to more advanced technologies and platforms is a current issue for many software organizations. Model-Driven Modernization combined with Software as a Service delivery model is a very promising approach, which possesses a lot of advantages, including reduced costs, automation of migration activities and reuse of system functionality. However, a drawback of such an innovative modernization approach is that it lacks mature software process models to guide its adoption. Thus, a methodology for seamless execution of different migration and deployment activities is quite needed. On the other hand, agile development methods have been successfully adopted in various projects, which partly or thoroughly use the engineering and delivery models exploited in the modernization process. This paper presents how a particular methodology for Model-Driven Modernization with deployment to the Cloud is enriched with agile techniques to address different challenging issues. The extended agile methodology could be used by organizations which have already applied agile software development as well as by organizations that plan to introduce it in their work.**

Keywords- Cloud computing; Agile Methodology; Model-driven Modernization; Software as a service

I. INTRODUCTION

Cloud computing and Service-Oriented Architecture (SOA) have recently been recognized as very promising approaches which provide cost-efficient and reliable services. Migration to the Service Cloud paradigm implies transformation of legacy software systems to SOA with deployment in the Cloud. Nowadays, the popularity of the Software as a Service (SaaS) cloud model is growing fast. The SaaS model reduces the infrastructure costs for customers and offers flexible license payment schemas. Despite its numerous advantages, building a SaaS system from scratch to replace the outdated software of an organization might not be a reasonable investment. A modernization approach based on reusing and integrating the company's legacy applications is a better solution.

Model Driven Modernization is a recent approach, whose aim is to provide automation of most of the migration activities and reuse of legacy functionality. OMG's Architecture Driven Modernization (ADM) provides support for MDM but is in its earliest stage. They envision a set of automated tools that can disassemble a legacy software system, transform the components in high-level models, reconfigure these models using the best-practices from Model Driven Architecture (MDA), and finally regenerate a modern system.

REMICS (REuse and Migration of legacy systems to Interoperable Cloud Services) is an EU FP7 research project with the objective of supporting the modernization of legacy systems to service cloud by providing a model-driven methodology and tools. REMICS proposes to improve existing approaches and extend them when needed to provide a holistic view of software migration that covers the whole process with a methodology, tools, languages and transformations. The methodology developed in the REMICS project covers the whole life cycle of the migration process. It proposes a traditional sequential approach to software engineering.

Agile methodologies, on the other hand, have been su

ccessfully applied in various contexts, including the ones related to the REMICS project. Thus, the question of whether agile software development can benefit the modernization process is quite adequate and particularly interesting. The aim of the study presented in the paper is to propose and describe a particular agile extension of the general REMICS methodology. A 5-step approach is used to specify how the traditional methodology can be enriched with appropriate agile techniques. The new agile methodology could be used by organizations, which have already applied agile software development as well as by organizations that wish to introduce agile methods in their work.

The rest of the paper is organized as follows. Section 2 presents in more details the REMICS project and related technologies, and thus describes the context of the modernization methodology. A current state-of-the-art of agile adoption in areas related to REMICS is also included. Information on the general REMICS methodology is provided in Section 3. Section 4 describes the approach followed in the creation of the agile extension. In section 5, the scrum types that are defined in the new agile modernization methodology are presented. Section 6 briefly describes how the applicability of the proposed agile REMICS methodology was studied. Finally, Section 7 concludes the paper.

 ISBN: 978-1-61208-280-6

II. BACKGROUND

The present section covers the background on which the agile extension for a migration methodology is created. On one hand, it is the REMICS project, which outlines the specific context of the modernization approach. On the other hand, it is the available research on agile adoption in areas related to this particular modernization approach.

A. The REMICS Project

The REMICS project promotes a new development paradigm for migration of legacy systems to the service cloud platforms through innovative model-driven technologies. The model-driven approach followed in the project is taking advantage of OMG's ADM (Architecture-Driven Modernization [1]), KDM (Knowledge Discovery Metamodel [2]), SoaML (Service-oriented architecture Modeling Language [3]) and UML profiles. The baseline concept is the ADM by OMG. In this concept the modernization starts with the extraction of the architecture of the legacy application. Having this architectural model facilitates the analysis of the source system, the identification of the best ways for its modernization and the incorporation of MDE technologies for generating the target system. The project intends to significantly enhance this generic process by proposing a set of advanced technologies for architecture recovery and migration, including innovative technologies such as Model-Driven Interoperability and Models@Runtime. Model-Driven Interoperability is a rather new domain, which builds on top of long history on data and service interoperability. Semi-automated methods that assist users to handle interoperability issues between services are also addressed in REMICS.

The REMICS project includes extending KDM and SoaML to cover concepts related to the migration to SOA and Cloud computing paradigms. Software as a Service (SaaS) is one of the delivery models of Cloud computing whose popularity and usage is growing rapidly in the recent years. The SaaS cloud model provides advantages for both software providers and users. The SaaS delivery model uses a multitenant architecture where a single application is delivered to millions of users through Internet browsers. The advantages offered for the SaaS end user is that installing software is avoided and complex software and hardware requirements can be rapidly fulfilled. In addition, end users do not require upfront licensing and can choose among different payment schemas. Organizations that offer software minimize their support costs and initial investments by outsourcing hardware and software provision and maintenance to the SaaS provider. The provider of SaaS software takes care of the security, availability and performance of the software because they are in charge of the deployment of the system.

A drawback of such a Modernization approach that uses so many innovative technologies is that it lacks mature software process models to guide their adoption. Thus, a methodology for seamless integration and execution of different migration and deployment activities is quite needed. A report on the state of the art on Service Cloud migration methodologies [4] showed that while there are several methodologies for developing service-oriented systems, they are mostly based on developing systems from scratch and not using a legacy system as basis for identifying and implementing services. On the other side, in the context of REMICS project, migration tools and methods need to be integrated with model-based development methods. The migration to interoperable cloud infrastructure introduces some challenges that are not addressed in the current methodologies in a complete way. The state of the art report showed that for the existing cloud computing platforms and cloud deployment model there is no methodology that guides developers through the process of selecting technology and migration to cloud. Additionally, while state of the art in SOA is quite established and covered in literature, the cloud design patterns are more an ad-hoc knowledge that still has to be studied. Existing approaches and methods for transforming a legacy system into a cloud compatible system have some shortcomings in the way they treat interoperability, reliability, scalability, configurability or multi-tenancy. Therefore, a comprehensive process model is needed, which helps organizations improve their technical know-how required for successful migration of their software.

B. Agile Adoption

Agile development has been on the cutting edge of both software industry and research for a decade. By shortening time-to-market and responding to the changing requirements of the business, agile approaches are reasonable alternative to the traditional waterfall development methods. The two most popular and widely adopted agile methods through the development community are Scrum [7] and Extreme Programming (XP) [8]. While Scrum provides a framework for managing and organizing agile projects, XP includes practices that are more technologically oriented and supports different development activities such as programming, code integration and testing. The two methods, as well as a hybrid between them, are used in more than two thirds of the agile projects surveyed by VersionOne in 2011 [9].

Nowadays, agile methods and techniques are being widely adapted and successfully applied in many business and application domains, where they are combined with a variety of technologies and platforms, including Model-Driven Development (MDD), Model-Driven Modernization (MDM), SOA and Cloud computing. The combination of MDD and agile software development (known as Agile Model-Driven Development (AMDD)), is the most broadly researched and used among the four technologies mentioned above. A review of existing AMDD methodologies is available by Matinnejad [10], Picek [11] and Mahé et al. [12]. Although there are many existing SOA methodologies, only few methodologies were found to be specifically concerned with incorporating agile software development. Some to mention here are:

- the Agile Service-Oriented Process (ASOP) presented in [13];
- the Xplus framework process model proposed by Shin and Kim [14];

 ISBN: 978-1-61208-280-6

- the combination of Rational Unified Process (RUP) to exploit SOA and a detailed development life cycle based on Agile Unified Process (AUP) [15];
- and the Continuous SCRUM [16].

The later uses a triple-sprint-overlap pattern and some additional best engineering practices in order to sustain a weekly release cycle of a SaaS and PaaS based software system. There are few methodologies, which combine agile software development with both MDD and SOA. The mScrum4SOSR is one such methodology, proposed by Chung et al [17]. The methodology extends Scrum with UML modeling and XP techniques in order to provide comprehensive service-oriented software reengineering process model. Another similar approach is the Model-driven Rapid Development Architecture (SMRDA) - an iterative development approach which unifies SOA and MDD in order to enhance the efficiency of the development efforts and the reusability of the developed services [18].

Based on the conducted review we could claim that there is no agile methodology in the literature which is specifically concerned with model-driven modernization with deployment on the service cloud. However, agile methods as Scrum and XP have been examined, combined and successfully applied in areas, closely related to the REMICS project.

III. GENERAL REMICS MIGRATION METHODOLOGY

The migration of software systems to cloud infrastructures can be viewed from two different perspectives - business and technology. The business dimension is focussed on the migration of the system to support the cloud business models and usually involves additional functional requirements in order to provide implementation of cloud related support activities such as billing and legalisation. The technology dimension is focussed on the migration of the system so it is able to run in the new cloud environment, taking advantage of the new technologies without adding too much additional functionality. It usually involves new non-functional requirements over the legacy system, as well as few functional requirements.

In general, REMICS migration methodology is focused mainly in the evolution of the technology model. There are 7 activity areas defined in the REMICS methodology, which cover the full life cycle of a legacy system Modernization to the Cloud. Fig. 1 depicts the main activity areas. Tools and techniques in REMICS project support all but the one of the activity areas. Only the withdrawal does not receive special support. For the purpose of the study presented in this paper, a brief description of the activity areas with the composing activities is presented in the sections below. They are the basis on which agile extension will be specified. More detailed information of the methodology is available in [19].

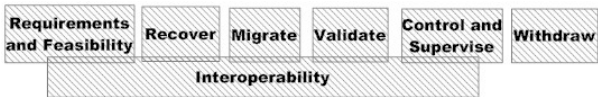


Figure 1. REMICS activity areas

A. Requirements and Feasibility Activity Area

In the *requirements and feasibility activity area* the migration requirements for the system are gathered and the main components of the solution and their implementation strategy are identified. The purpose is not an exhaustive description of all requirements of the source system, but the description of the requirements that will require development effort and will be used as a basis for the validation of the system. Feasibility analysis is needed since in a migrated system not all the parts are equally reusable. In some cases the best way to reuse a component may be to wrap it, in other cases to reengineer, or to replace it with an external one, or to implement it from scratch, etc.

The main activities that are included in the *requirements and feasibility activity area* are the following:

- Describe the system;
- Apply techniques to evaluate feasibility;
- Identify actors;
- Identify additional requirements and specify new requirements in UML diagrams;
- Establish validation criteria; and
- Elaborate glossary.

B. Recover Activity Area

The purpose of the *recover activity area* is the recovery of the knowledge from those legacy components that during the feasibility analysis has been pointed as candidates to be reengineered. The use of recover methods and tools will provide the application model of the legacy system as well as information on the requirements and the testing procedures for the migrated code. The system knowledge is recovered in KDM format from which UML system models and requirements specification in Requirement Specification Language (RSL) are generated. The following activities are part of the *recover activity area*:

- Collect the code;
- Recover system knowledge;
- Refine system knowledge;
- Generate system model;
- Generate system requirement; and
- Recover application testing.

C. Migrate Activity Area

In the *migrate activity area* the target system is defined and implemented using the elements identified during the requirement and recover phases. This includes also the design and implementation of the components necessary for the SaaS application and the development of the service-oriented architecture. The component SOA model is used to define the deployment model of the system. The deployment model contains the identification of the location of the different components distributed in the cloud. When defining this deployment model, it is also necessary to take into account possible constrains (organizational, legal, etc).

The activities that are included in *the migrate activity area* are:

- Complete definition of system requirement;
- Definition of service architecture;

 ISBN: 978-1-61208-280-6

- Definition of cloud architecture;
- Implementation design;
- Generate code stubs of the system; and
- Complete the code.

D. Validation Activity Area

The purpose of the *validation activity area* is to define testing strategy to verify that the migrated system implements the requirements identified and that the components (including those not reengineered) and services work properly. This validation phase includes not only functional validation but what is more important, non-functional validation, especially performance, reliability and security. In the case of cloud computing applications these three aspects must be stressed on. Testing procedures are defined depending on the specifics of the application and particular system requirements.

The following activities are part of the *validation activity area*:

- Define testing infrastructure;
- Identify and refine requirements to be tested;
- Generate acceptance testing;
- Import models elements to be tested;
- Define testing procedures; and
- Implement testing strategy.

E. Supervise Activity Area

The *supervise activity area* provides elements to monitor and control the performance of the system when deployed in the Cloud and to modify that performance. A company can monitor constantly the performance of the application once it has been provisioned as a service, so it can be improved in performance, reliability and resources used. As well, the system can be supervised of possible degradations.

Activities included in the supervision process are:

- Identify monitoring procedures;
- Implement monitoring procedures;
- Monitor the performance of the system in the cloud;
- Detect deviations;
- Perform corrective actions; and
- Monitor corrective actions.

F. Interoperability Activity Area

The *interoperability activity area* provides tools that solve interoperability problems with 3rd part providers or any external components and services. Interoperability is a crosscutting activity to the general methodology that deals with the interoperability issues that affect SaaS along the other activity areas.

Activities for performing interoperability are the following:

- Identify interoperability problems/scenarios;
- Define interoperability related requirements;
- Perform interoperability analysis;
- Implement interoperability components ; and
- Interoperability monitoring.

IV. AGILE EXTENSION TO THE MIGRATION METHODOLOGY

The extension of the general REMICS migration methodology with agile techniques follows a 7-steps approach – identify, analyze, select, define, formalize, evaluate and adapt. The first 5 steps, in which an initial agile methodology is defined, are described in details in the following subsections. The applicability of the initial methodology is studied in the evaluation step and is presented in the following section. As a last step of the approach, further enhancements of the initial methodology are suggested based on the output of the evaluation step. The adaptation step is subject to future research.

A. Identify and Analyze

The identification and analysis of agile methods and techniques in the context of our study was conducted in two consecutive phases. During the first phase, the challenges of all related fields, incl. MDD, SOA, Cloud computing and Software modernization, were extracted, analyzed and synthesized through a systematic literature review [20]. In the second phase various agile techniques were evaluated through the Delphi method in terms of their potential to overcome the extracted challenges.

The systematic review covered total of 84 articles, which were either describing the current state of research and practice in any of the above mentioned four related fields or were identifying and discussing the challenges these areas possess to both academia and industry. The full texts of these articles were thoroughly examined in order to extract all relevant challenges. The extracted challenges were further consolidated into total of 52 challenges and sorted into two categories – organizational and technical challenges. Organizational challenges included process-oriented and people-oriented challenges from all levels of the organization (e.g. competence acquisition, process reengineering, addressing external dependencies, etc.), while technical challenges included design, implementation, verification and deployment challenges, and thus were mostly product and technology-oriented. Among the most cited organizational challenges in the MDD field are the lack of mature tools, integrated development environments and off-the-shelf infrastructure, competence acquisition and reliance on high level models. Popular technical challenges for SOA applications are service design, addressing security, interoperability and other quality aspects and testing services. In the Cloud computing field, among the most cited challenges are trust, security and privacy, external dependencies and vendor lock-ins. Common technical challenges faced during software modernization are ensuring behavioral equivalence and extracting business and technical knowledge from legacy systems. A thorough analysis of the extracted challenges and their implication to agile software development could be found in our previous works [21-22].

During the second phase, various agile techniques were evaluated for their potential to address the challenges, extracted by the review process. These agile techniques were taken from Scrum and XP agile development methods. The

 ISBN: 978-1-61208-280-6

methodology used to evaluate these techniques was the Delphi method. More specifically, the Pfeiffer's three step process was followed [23]. During the recommendation step, a panel of experts (with an average of 9 years of both academic and industrial experience in agile software development) was asked to review the list of extracted challenges and recommend agile techniques based on the challenges they could address. In the evaluation step, a consolidated list of recommended agile techniques was sent to each expert to further evaluate the relevance (on a five-point rating scale) of all techniques in regard to all extracted challenges. Finally, during the consensus phase, the consolidated list, together with the experts' ratings was sent once again in order to discuss big differences in ratings and gain consensus. In result, a sorted list of recommended agile techniques was created based on the number of challenges they could address. This list is shown in Table 1.

As seen from Table 1, the agile techniques with the highest rating were Small releases, Planning game and Whole team from XP, and Sprints, Cross-functional teams and Sprint planning meeting from Scrum. Among the arguments for this were receiving feedback quickly, increasing responsiveness to change, building trust and confidence, enhanced collaboration, clarification of team responsibilities and service ownership, early escalation of quality concerns, effective acquisition of competencies and expertise, early delivery of customer value, and many more [21-22]. Other agile techniques, which were also highly recommended by the experts, were Pair programming and Continuous integration from XP, and Product and Sprint backlogs.

TABLE I. EVALUATION OF AGILE TECHNIQUES BASED ON THE CHALLENGES THEY ARE EXPECTED TO ADDRESS

Agile technique	Number of addressed challenges
Extreme programming (XP)	
Small releases	38
Planning game	29
Whole team	29
Pair programming	26
Continuous integration	23
Test-driven development	21
System metaphor	14
Collective code ownership	14
Refactoring	9
Simple design	7
Coding standards	6
Scrum	
Sprint	38
Cross-functional teams	35
Sprint planning meeting	33
Product backlog	26
Sprint backlog	26
Product owner	17
Daily scrum	16
Scrum master	13
Scrum of scrums	11
Sprint review meeting	11
Sprint retrospective	7
Sprint burn down chart	4

B. Select

During the selection step, a set of techniques to be included in the initial methodology is identified. As well, techniques that are a subject to modification are chosen. The selection of the initial set of techniques is guided by the principle of incremental methodology design [24], according to which easy to apply techniques are included in the beginning while more challenging techniques are added iteratively. The other two criteria, which are used to select agile techniques for the initial methodology, are taken from the context of the study. The first one is the rating of the agile technique from the sorted list created in the previous step. Some techniques were selected because they were suggested to improve the general REMICS methodology by the industry partners who have been involved in different project case studies. As a result, fourteen techniques out of twenty three are included in the initial agile methodology.

C. Define

During the define step, the new agile methodology is specified. As well, the techniques which are modified to suit the requirements of the migration process and the new ones are described. The agile REMICS methodology proposes a particular implementation of Scrum methodology for large projects. It is based on the so-called Type-C SCRUM [25] in which a number of integrated overlapping scrums are executed. The methodology has been adjusted for the characteristics of the REMICS project by defining a number of Modernization Scrum types. For each of these types, the main Scrum techniques have been modified in the following way:

- Modernization sprints – a number of Sprint types with particular activities depending on REMICS activity areas;
- Modernization product backlogs – each sprint type has a corresponding product backlog. The *Product owner* manages different Product backlog types; and
- Modernization teams - different *Scrum team types are* associated with each Sprint type depending on the required skills for particular sprint type. Teams are self-organizing but with a certain degree of specialization (which is in contrast to general scrum teams) due to the diverse and not so common skills needed for activities in the Sprint types. The Scrum team types integrate the *Whole team* technique of XP as well. Teams are (preferably) collocated and a business expert or a customer is part of the team.

In addition to the Scrum types, there are a couple of agile development techniques that have been adjusted for the extensively used model-driven engineering activities in the REMICS project:

- Modeling by two - based on the *pair programming technique* of XP the modeling by two is done by two people with different roles who work simultaneously on one model in order to exchange, analyze and synthesize knowledge in models better, faster and more easily

 ISBN: 978-1-61208-280-6

- Pair modeling - two analysts work together on a model;
- Collective model ownership - created models are continuously integrated in a common code base; and
- Continuous modeling - models are collectively owned by the whole team.

To make the agile REMICS methodology more effective, a new technique called Shifting team member has been added to the set of agile techniques. A team member moves among teams in different sprint types. When a new sprint begins, (s)he shifts to another team where the team member plays the same or similar role.

D. Formalize

The methodology is formally specified using the Eclipse Process Framework (EPF). EPF [26] is an open source solution that implements the SPEM modelling language. SPEM (Software process engineering meta-model) [27] is a specification from the OMG that addresses the standardised definition of software development methodologies. Other tools can latterly automatically process such specifications for different purposes. The agile methodology is based on the general REMICS methodology so they are specified in different packages in which activities refer to each other.

In addition to EPF, the agile REMICS methodology will be implemented in the language proposed in the OMG FACESEM RFP by the SEMAT working group [28]. Both options for specification of REMICS methodology will be compared and the end of the project will give the recommendations for using each one.

V. SCRUM TYPES

The section describes the five scrum types of the proposed initial agile REMICS methodology. To a great extend they conform to the activity areas of the general REMICS methodology. The interoperability activities are added to the related scrum types. Fig. 2 presents the basic components of the scrum types. The activities from the activity areas of the initial REMICS methodology are present in each scrum type. They are modified appropriately to introduce different agile techniques e.g. modelling by two technique is applied in refine system knowledge activity during the recovery scrum. A set of new activities to support agile techniques, such as sprint backlogs and sprint retrospective, are added in each scrum type. Activities that don't support iterative execution in sprints and serve as preconditions for execution of other activities are gathered in the so called initiation or initialization activity. The initiation activity is executed just ones in the beginning of particular scrum, while the initialization activity might be performed several times in a scrum e.g. each time a new component is to be recovered or migrated.

The presentation here outlines only the major activities of the respective sprint types and how the identified agile techniques are applied. A detailed information of the methodology with activity inputs and outputs, modernization team roles and support materials is available in a final REMICS project report [29]. The last subsection discusses possible life cycles of a Modernization project in which the five scrum types are executed.

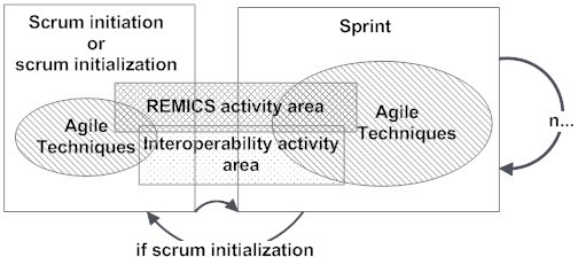


Figure 2. Scrum type

Each project scrum starts with Initiation project activity, which consists of four activities that are performed as sessions and meetings with the whole project team:

- Goals of migration;
- Describe the system;
- Apply techniques to evaluate feasibility;
- Identify actors and initial scrum teams; and
- Prepare the project product backlog and prioritise components for implementation;

The output of the activity is the project product backlog with prioritized components and implementation strategies.

A. Requirements scrum

The requirements scrum contains activities for new requirements identification and specification. As well, the scrum handles requirements that address interoperability issues. There are two main activities in the requirements scrum:

- Requirements scrum initiation; and
- Requirements sprints.

The requirements scrum initiation activity is held in the beginning of the scrum and has two activities which are new with regard to the general REMICS methodology:

- Prepare and demo product backlog for requirements scrum; and
- Define general deployment model.

Requirements sprints are one or more sprints executed after the initiation activity. Each requirements sprint starts with a planning meeting and ends with a retrospective. As well, it contains related activities from the general *Requirements and Feasibility activity area* and interoperability requirements identification and specification. The Modelling by two technique is applied for requirements identification and specification when an analyst and a business expert work together on new requirements specification using UML.

B. Recovery scrum

The recovery scrum contains activities for recovering of the system requirements from the existing code of the legacy application. There are two main activities in the recovery scrum:

- Recovery initialization; and
- Recovery sprints.

 ISBN: 978-1-61208-280-6

The recovery initialization activity is held every time when the recovery of a particular component starts. The initialization contains three activities:

- Collect the code;
- Recover system knowledge; and
- Prepare and demo product backlog for the component which is to be recovered.

There are a number of recovery sprints that are carried after the recovery initialization. The sprints start with a sprint planning meeting. At the end of the sprints, a retrospective meeting and a demo are performed. The activities in the recovery sprints are executed iteratively and contain refinement of system knowledge, generation of system models and requirements in RSL and recovery of application testing. To support the iterative execution, the technique of continuous modelling and collective model ownership are applied. The system knowledge refinement and system modeling are performed together with a business expert and a developer.

C. Migration scrum

The migration scrum contains activities for development of all the system components depending on their implementation strategies – wrapped, recovered and new components. Migration starts when a specification of a component is ready- either of a new component or a recovered one.

There are two main activities in the migration scrum:

- Migration initialization; and
- Migration sprints.

The migration initialization activity is executed in the beginning of each new component that is to be migrated. Migration initialization contains three activities:

- Prepare and demo product backlog for migration scrum;
- Componentization of UML or RSL models;
- Definition of overall cloud architecture.

One or more migration sprints are executed after the initialization activity. Migration sprints contain the activities of the general REMICS migration activity area, interoperability analysis and interoperability components implementation. As during the migration there are a couple of activities, involving knowledge from various innovative and complex areas (e.g. definition of service and cloud architectures), a pairing with a team member technique is suggested. The common for all sprints planning and retrospective activities are also included in the migration sprints. Testing activities of an isolated part of the system is added in the migration sprint so that the demonstrated increment of the system is verified.

D. Integration and Validation Scrum

During the integration and validation scrum, the migrated parts of the system are integrated and validation activities of the new functionality are performed. Regression testing is also part of the validation activities. There are two main activities in the integration and validation scrum:

- Integration and validation scrum initialization; and
- Integration and validation sprints.

The initialization activity is performed in the beginning of the integration and validation scrum and each time a part of a new component is migrated. Integration and validation scrum initialization contains three activities:

- Define testing infrastructure;
- Identify and refine requirements to be tested; and
- Create and demo product backlog for the scrum.

After the initialization activity, a number of integration and validation sprints for a particular component are executed. The activities from the general validation activity area are included in the integration and validation sprints. A new integration activity is added since the validation is performed incrementally.

E. Control and Supervise Scrum

The control and supervise scrum contains activities for monitoring of the deployed system. It starts when the first release of the system is deployed in the Cloud. There are two main activities in the control and supervise scrum:

- Control and supervise initialization; and
- Control and supervise sprint.

The initialization activity is performed every time a new release is deployed since the monitoring procedures may vary from one release to another. Control and supervise initialization contains two activities:

- Identify monitoring procedures; and
- Prepare the control and supervise product backlog.

After the initialization, there might be one or more sprints to monitor the running system, detect deviations and correct them. When a deviation is identified, it is added to the product backlog to be corrected in some of the following sprints. The control and supervise sprints contain the activities from the corresponding activity area in the general REMICS methodology as well as monitoring of interoperability.

F. Life Cycle

A possible life cycle of a modernization project in which the five scrum types are executed is shown on Figure 7. The scrums are overlapping and executed in parallel when possible. The main point is to have releases as soon as possible and to provide continuous and early feedback by a number of sprints. The figure shows one particular lifecycle, however, depending on the project size and staffing possibilities there might be different arrangements of the scrums and sprints. For example, the scrums might be executed successively instead of simultaneously as shown in the figure. As well, the sprints inside each scrum can be executed in parallel, if there are more than one teams of particular type dedicated to the project.

There are forward relations between the scrums as well as backward relations e.g. if during a migration sprint new requirements emerge they will be part of some of the subsequent requirements sprints; problems found during the integration and validation sprint will be handled during the new migration sprints, etc. The project ends with a sprint of the control and supervise scrum.

 ISBN: 978-1-61208-280-6

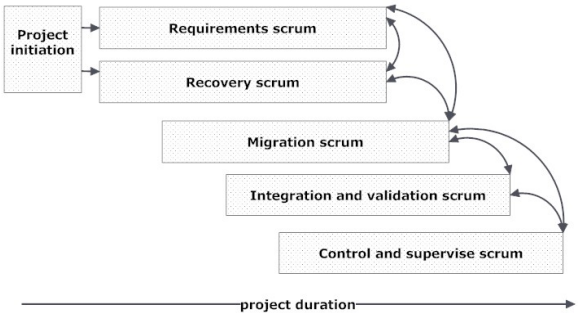


Figure 3. Project life cycle with Modernization scrums

VI. APPLICATION

As part of the evaluation step, the applicability of the proposed agile REMICS methodology was studied in two ways. First, a questionnaire was sent to all of the four case study providers participating in the project. The four case studies differ in business domains and technologies used in their legacy systems. A representative of each case study was asked to answer questions on whether particular agile technique is applicable to his/her case study and to what extent. In addition, for each technique it was studied why it is not applicable and what problems it could address if applied. The information gathered in the questionnaire is supposed to give insight on characteristics of each project as well as specifics of the general modernization process that affect application of the agile REMICS methodology. All the responders has stated moderately or very familiar with both Scrum and XP methods so their judgment could be considered adequate. The overall results showed that most of the suggested agile techniques are applicable in the four case study projects. There were no major issues, considered by the case study providers, which could prevent the proposed agile techniques from applying in their projects.

As a second way for evaluation of the methodology, it was applied in one of the case study projects. Based on the responses of the questionnaire of agile practices applicability and further interviews with the providers of the case study, a customized agile methodology was specified and deployed in the project. Since a pilot functionality of the legacy system was chosen for migration, there was only one team who executed all of the scrums. The agile REMICS methodology was applied for three months with two-week-long sprints. Currently, the results of the application are studied. Satisfaction of the applied techniques is surveyed and suggestions for improvements are gathered. Preliminary results showed that most of the agile techniques were successfully applied, but some adjustments were needed. Based on the feedback provided by case study provider the proposed agile methodology will be further improved.

VII. CONCLUSION

Some of the benefits of using agile methods in software development projects are decreased time-to-market, minimized risk of project failure and growing confidence and satisfaction of project stakeholders. In the present years, the industry is facing urgent need for modernization of outdated software system. Along with that, the popularity of SaaS application is growing fast. The SaaS cloud model provides advantages for both software providers and users. By introducing agile approaches to the development of SaaS applications their numerous advantages could be beneficial for software organizations.

In the current paper, an approach for extension of a particular methodology for model-driven migration of legacy systems to the Service Cloud is proposed. The steps of the suggested approach describe how agile practices and techniques have been identified, analyzed, selected, defined and evaluated to enrich the REMICS project methodology with appropriate techniques. The agile extension of the methodology further addresses challenges of the modernization process and provides support for organizations to move their legacy systems to SaaS applications following the agile development principles. The new methodology is evaluated in industry case studies in two ways. Firstly, a questionnaire of the applicability of agile practices was conducted with the providers of four case studies. Their expert opinion was considered to evaluate applicability of the methodology in different migration project settings. Secondly, the agile REMICS methodology was applied in one case study for three months. Currently, the results of the application are analyzed. Preliminary results showed that most of the agile techniques were successfully applied, but some adjustments need to be done. As a last step of the proposed approach for extension of REMICS modernisation methodology, the proposed agile methodology will be further improved based on the feedback provided by case study provider.

ACKNOWLEDGMENT

The research leading to these results has been developed in the context of the REMICS project (www.remics.eu) partly founded from the European Community's Seventh Framework Programme under grant agreement n° 257793. The authors wish to acknowledge the Commission for their support. We also wish to acknowledge our gratitude and appreciation to all the REMICS Project partners for their contribution during the development of various ideas and concepts presented in this paper.

REFERENCES

[1] ADM. OMG Architecture-Driven Modernization. Available: http://www.omgwiki.org/admtf/doku.php, [retrieved: 05, 2013]

[2] KDM. OMG ADM Knowledge Discovery Metamodel. Available: http://www.omg.org/spec/KDM/, [retrieved: 05, 2013]

[3] SoaML. OMG Service-Oriented Architecture Modeling Language. Available: http://www.omg.org/spec/SoaML/1.0.1/PDF/, [retrieved: 05, 2013]

[4] REMICS Project deliverable: State of the art on modernization methodologies, methods and tools. Available: http://www.remics.eu/system/files/REMICS_D2.1_V1.0_LowResolution.pdf, [retrieved: 05, 2013]

 ISBN: 978-1-61208-280-6

[5] S. Ambler, “Agile Software Development at Scale Balancing Agility and Formalism in Software Engineering.” vol. 5082, B. Meyer, et al., Eds., ed: Springer Berlin / Heidelberg, 2008, pp. 1-12.

[6] S. W. Ambler, The Object Primer: Agile Model-Driven Development with UML 2.0: Cambridge University Press, 2004.

[7] K. Beck, Extreme Programming Explained: Embrace Change, Addison_Wesley Professional, 1999

[8] K. Schwaber and M. Beedle, Agile Software Development with Scrum, Prentice Hall, 2002

[9] VersionOne. (2011, State of Agile Development Survey Results. Available: http://www.versionone.com/ state_of_agile_development_survey/11/, [retrieved: 05, 2013]

[10] R. Matinnejad, “Agile Model Driven Development: An Intelligent Compromise,” in Software Engineering Research, Management and Applications (SERA), 2011 9th International Conference on, 2011, pp. 197-202.

[11] R. Picek, “Suitability of Modern Software Development Methodologies for Model Driven Development,” Journal of Information and Organizational Sciences, vol. 33, pp. 285-295, 2009.

[12] V. Mahé, B. Combemale, and J. Cadavid, “Crossing Model Driven Engineering and Agility: Preliminary Thought on Benefits and Challenges,” in 3rd Workshop on Model-Driven Tool & Process Integration, in conjunction with Sxth European Conference on Modelling Foundations and Applications, 2010, pp. 97-108.

[13] A. Qumer and B. Henderson-Sellers, “ASOP: An Agile Service-Oriented Process,” Proc. Software Methodologies, Tools and Techniques 07, 2007, pp. 83-92.

[14] S. W. Shin and Haeng Kon Kim, “A Framework for SOA-Based Application on Agile of Small and Medium Enterprise,” in Computer and Information Science , Roger Lee and H. Kim. Eds., Springer Berlin Heidelberg, 2008, pp. 107-120

[15] I. Christou, S. Ponis and E. Palaiologou, “Using the Agile Unified Process in Banking,” IEEE Softw., vol. 27, 2010, pp. 72-79.

[16] P. Agarwal, “Continuous SCRUM: agile management of SAAS products,” Proc. of the 4th India Software Engineering Conference, Thiruvananthapuram, Kerala, India, 2011, pp. 51-60

[17] S. Chung, D. Won, S. Baeg and S. Park, “A Model-Driven Scrum Process for Service-Oriented Software Reengineering: mScrum4SOSR,” in The 2nd International Conference on Computer Science and its Applications (CSA 2009), Jeju Island, Korea, 2009, pp. 1-8.

[18] B. Wang, C. Wen and J. Sheng, “A SOA based Model driven Rapid Development Architecture - SMRDA,” Proc. of the 2nd International Conference on Education Technology and Computer (ICETC 2010), Shanghai, China, 2010, pp. 421-425

[19] REMICS. (2011, Project deliverable: REMICS Methodology. Available: http://www.remics.eu/system/files/REMICS_D2.2_V1.0.pdf, [retrieved: 05, 2013]

[20] P. Brereton, B. A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, “Lessons from applying the systematic literature review process within the software engineering domain,” J. Syst. Softw., vol. 80, 2007, pp. 571-583.

[21] S. Stavru, I. Krasteva and S. Ilieva, “Challenges for Migrating to the Service Cloud Paradigm: An Agile Perspective,” Web Information Systems Engineering – WISE 2011 and 2012 Workshops, Springer Berlin Heidelberg, 2012, pp 77-91.

[22] S. Stavru, I. Krasteva and S. Ilieva., “Challenges of Model-Driven Software Modernization: An Agile Perspective,” Proc. The 1st International Conference on Model-Driven Engineering and Software Development (MODELSWARD 2013), Barcelona, Spain, 2013.

[23] J. Pfeiffer, New look at education: systems analysis in our schools and colleges: Odyssey Press, 1968.

[24] K. Beck and C. Andres, Extreme Programming Explained: Embrace Change (2nd Edition): Addison-Wesley Professional, 2004.

[25] J. Sutherland, “Future of Scrum: Parallel Pipelining of Sprints in Complex Projects,” presented at the Proceedings of the Agile Development Conference, 2005.

[26] EPF. Eclipse Process Framework Project (EPF). Available: http://www.eclipse.org/epf/, [retrieved: 05, 2013]

[27] SPEM. OMG Software Process Engineering Meta-model 2.0. Available: http://www.omg.org/spec/SPEM/2.0/PDF/, [retrieved: 05, 2013]

[28] SEMAT working group, Available: http://semat.org/, [retrieved: 05, 2013]

[29] REMICS. Project deliverable: REMICS Methodology with Agile Extension. Available: http://www.remics.eu/ publicdeliverables [retrieved: 05, 2013]

 ISBN: 978-1-61208-280-6

Ubiquitous System to Enhance the Supply Chain

Cándido Caballero-Gil, Jezabel Molina-Gil, Pino Caballero-Gil
Department of Statistics, Operations Research and Computing
University of La Laguna
Tenerife, Spain
Email: {ccabgil,jmmolina,pcaballe}@ull.es

Alexis Quesada-Arencibia
Institute of Cybernetic Science and Technology
University of Las Palmas de G.C.
Las Palmas, Spain
Email: aquesada@dis.ulpgc.es

***Abstract*—Transportation and logistics include delivery, movement and collection of goods through roads, and in the international case also through ports and airports. Consequently, they usually involve many different actors, what complicates management, reducing efficiency and effectiveness. In particular, time, boundaries, and interdependencies are the main difficulties in any supply chain. Besides, several security challenges are raised due to unintentional errors or intentional attacks. Existing technology such as RFID, with its potential to automate product authentications, makes possible to solve, or at least to reduce, most of the possible negative effects caused by the mismanagement of the supply chain process. In this line, technology included in the Internet of Things (IoT) can be used to enhance several aspects of the supply chain management, helping to improve demand management, customization, and automatic replenishment of out-of-stock goods while reducing inventory and distribution costs, as well as counterfeit versions of name-brand items. It also allows creating new safe and efficient schemes that help enterprises and organizations to improve quality of service and traceability to the management of transported goods. This paper proposes the IoT integration in the transport of merchandise, allowing its follow-up from the cloud. Preliminary research results indicate that the combination of ubiquitous technologies provides a more complete and efficient service.**

***Keywords*-Ubiquitous systems; Internet of Things; wireless technologies; transportation and logistics services.**

I. INTRODUCTION

Robust and real-time information is essential for the operation of any organization dedicated to the transportation and logistics sector. In particular, tracking and tracing goods is a process that can cover the determination of the current and past locations of items, report of the arrival or departure of objects and record of the identification of objects, and the location, time, and status where they were observed. Such pieces of information can be very helpful both internally and for customers. On the one hand, it can help to design and manage the Supply Chain (SC), adding value to the products, and establishing an additional customer service. On the other hand, internally, it allows examining how customers can get benefit by improving the use of resources in order to reduce storage costs, risk of loss or theft, etc.

Changes in customers' likes have increased the attention on the performance, design, and analysis of the SC in order to improve its efficiency and effectiveness, and to launch new products faster at lower costs. The managers of enterprises know that competitiveness is not only achieved by optimizing the manufacturing lines, but it is also important to improve the SC, enhancing the productivity growth.

According to the RAPEX report [1], the traceability of consumer products would improve Europe's ability to fight fraud and take action against unsafe products. This would be possible through a comprehensive control of the goods that reaches countries borders, including details such as date and place of entry as well as the origin and destination of the imports. It also stresses the importance of monitoring and control of goods at any time and place, since this would decrease the effects of dangerous goods that could enter the borders or thefts that may arise.

A study from 2004 [2] estimates that in the U.S. there were excess inventories of more than $117 billion and many enterprises lost $83 billion due to problems of coordination between different elements of the chain. In order to have an efficient SC, it is necessary to provide fast supply, be responsive to customer demand shifts, and alert about supply disruption.

In the transportation and logistics industry, automation in product monitoring and control, inventory, customer relationship management, fleet tracking, etc., is a typical issue dealt by the enterprises that offer solutions for the individual problems. The main goal of this work is to improve such solutions by making use of the IoT, through managing ubiquitous information about the transported goods with different types of communications and devices [3]. This work describes an innovative solution for the management of the complete SC process, which makes use of many different IoT technologies such as RFID, EPC, Wi-Fi, GPS, QR codes, etc. in a secure and efficient way. First, it provides the possibility of comparing the transported goods with the delivery note. Besides, it allows to observe, trace and check the merchandise from the source to the destination. In addition to this, the system offers an interface for fast checking of merchandise for authorities. Such information will benefit not only the authorities but also the exporters and importers, who can control their merchandise, ensure its reliability, optimize its transportation through adaptive travel route assignment, and provide added value to customers.

 ISBN: 978-1-61208-280-6

This paper is organized as follows. A brief survey of related work is given in Section II. Some preliminaries about security aspects and risk factors in the SC are included in Section III. Section IV presents the new ubiquitous solution. Finally, Section V gives conclusions and future works.

II. Related Work

The application of IoT in the transportation and logistics sector and the SC management [4] [5], requires dealing with many different problems related to reliability and security. In [6] the author classifies several existing models to improve the SC and describes different approaches for every model.

On the other hand, it is known that most security problems in the SC [7] are related to physical issues happening along the route, such as breaks in fencing, long stops at night, stops near residential areas, regular scheduled stops, unlocked containers and unmanned trains.

The main goal of this work is to improve the productiveness and security of the SC by using today's technologies. In this regard, [8] performs an analysis of the efficiency and effectiveness after applying different strategies to improve the SC, and [9] includes a complete review of more than two hundred papers about green SC management.

In this paper we also analyse an inherent problem of the used technology, which is the cloning problem [10], consisting in tampering goods. In order to prevent it, we propose the use of an authentication P2P solution based on Zero-Knowledge-Proofs, similar to the one presented in [11].

Solutions sharing specific characteristics with the proposed app can be found in the Google Play Store, which facilitate the tracking in the SC [12] [13]. However, the app proposed in this paper allows tracking packages from many carriers with a phone or tablet, starting form the delivery note, and tracing it in a secure way till the destination, combining different technologies.

III. Preliminaries

Using a combination of international security standards, industry best practice, regional legislative documentation and experience, solutions could be provided that would enhance security in the SC, as well as provide increased visibility and efficiency, by preventing and limiting the amount of cargo theft. Other recommendations to secure SC systems are to provide all logistics partners with a set of minimal security requirements, seal containers with highly secure locks, wrap all products with non-descriptive plastic in order to confuse potential criminals, and implement information security system to ensure that product information and transportation routes are not accessible to unauthorized personnel.

All the aforementioned recommendations may be seen as theoretical because they have sense only if they are applied in practice. Thus, this paper proposes a practical system using existing technologies, which allows enterprises, customers and organizations to have more robust information of the SC in real time, and to set up alarms in case of changes in goods containers in order to reduce risk in the SC.

Measurement is the first step that allows facing to the control and improvement of the SC. It is known that improvements in quality lead to lower costs and higher productivity because they result in less rework, fewer mistakes, fewer delays, and better use of time and materials. Therefore, measuring the performance allows to set goals and evaluate progress, by identifying key aspects to refine, analyzing the results of the refinements, and solving possible problems. In particular, in the SC, to develop an improvement of any area, it is important to know the starting situation, define a realistic goal, determine how to measure the progress in the system and develop an action plan. This has been the procedure followed when designing the proposal here presented.

The Japanese philosophy known as Just In Time (JIT) is a production strategy that strives to improve a business return on investment by reducing in-process inventory and associated carrying costs. To meet JIT objectives, the process has to rely on signals between different points in the process, which tell production when to make the next part. Implemented correctly, JIT focuses on continuous improvement, so our proposal can be a useful tool to reach that goal.

Enterprises without centralized SC governance can negatively impact procurement, manufacturing and time to market processes in SC, which can impact company's financial strategy. Security risk management is an essential part of the SC governance system to ensure that risks are identified in the entire value chain and mitigated to deliver financial goals. In particular, the stages of the SC that must be covered by the system are manufacturing and distribution, suppliers, transportation, retailers, central warehouse, docks, cranes, boats and wholesale and retail distribution center.

Different goals can be defined to improve the SC logistics. Some of them are to reduce costs and maximize profits, improve reliability, minimize inventory, reduce delivery time, maximize equipment utilization, increase flexibility, improving simulation, reduce work in process, and reuse. Practical tools to process the building of a dynamic SC model allow to have valuable insights and analyse the behaviour and characteristics of a SC. Most existing models and tools have been developed to address particular issues that can be classified into the following five categories:

- Optimization: Finding the optimal operational guidelines that maximize or minimize a factor, such as minimizing costs and/or risks and maximizing profits.
- Decision Analysis: Typically involves the quantitative evaluation and comparison of two or more alternatives.
- Diagnostic Evaluation: It is usually conducted when the cause of a particular problem is unknown.
- Risk Management: SC dynamics can be severely impacted by unanticipated disruptive events.
- Project Planning: Changes in SC parts can produce disruptions, and short-term/long-term inefficiencies.

 ISBN: 978-1-61208-280-6

IV. Proposed Ubiquitous System

This paper presents a secure and ubiquitous system to control the goods from their manufacture until their delivery to the end customer, which makes the work easier for custom authorities and all people responsible for goods in transit.

In particular, the new system allows checking whether the collected goods are correct by detecting any error in the delivery process, provides on-line checking mechanisms and keeps a full history of the goods transportation. All this is done taking into account the security needs required during the SC process because cryptographic algorithms are used to detect counterfeit information and to avoid unauthorized reading, writing or modification of labelling goods.

The proposal implies a minimum cost as it is based on affordable and usual devices such as RFID and smartphones, so not only minimizes economic costs by using cheap passive tags, but also reduces the effort to learn to use new technology because people are familiar with smartphones.

RFID technology is extended in enterprises related to SC and it allows us validate products easily and locate them [14] [15] fastly inside the container.

A typical SC consists of five actors: supplier, manufacturer, distributor, retailer, and customer. The proposed system has different parts according to the five steps in the SC where the goods may be: generation and extraction of QR data of container goods, RFID validation of goods, web service for fleet tracking and traceability, and Wi-Fi P2P request for customs check.

There are different kind of relationships in the SC and each one has different characteristics that must be resolved in a different way. Fig. 1 shows different technologies we proposed in order to have the best solutions for these relationships.

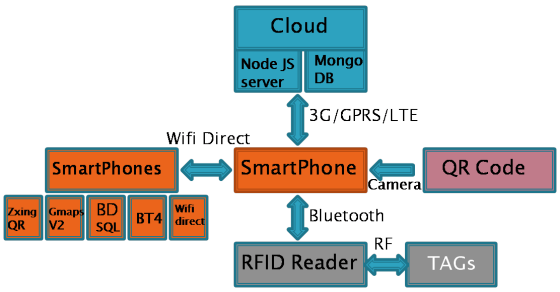


Figure 1. Related IoT technologies for the SC system

A. QR Container Receipt

The first step in the operation of the system is the generation of the container receipt describing relevant information about the products available in the container. The specific format of such a receipt consists of: receipt ID, 13-character codes identifying the products and number of each product, origin and destination places and corresponding dates, and other important data such as related enterprises and carriers.

The system generates a QR code containing all the information of the container receipt, and encrypts it so that the only way to read its content is by using the shared secret key. Such a QR code is printed in the same data sheet of the container goods.

The QR code with all information is generated with an on-line service in [16], to encrypt the information, data must be ciphered before generate the QR code.

B. QR Data Extraction

The driver who transports the goods executes the second step of the process. He/she has to use his/her smartphone to read the QR code with the app (see Step 1 in Fig. 2), and the content is decrypted with the shared key. After reading the code, the driver gets all the information of the QR container receipt.

The implementation in the Android platform for reading QR code is by using a library called Zxing, after this, data are saved in local device in a sql database. Information about origin and destinations are also shown in the device interface by using Google Maps Android API v2.

C. RFID Validation

With the list of transported products in the smartphone, the device may be used to read the RFID tags in the container through an RFID reader and a Wi-Fi interface to communicate with it. After this reading, the smartphone checks if every product in the list is in the container, (see Step 2 in Fig. 2). This is especially interesting not only in the loading of the goods in the container, but also in every delivery of goods in order to avoid possible errors. RFID technology is also useful to know the position where a product is inside the containter.

D. Web Service for Fleet Tracking and Traceability

On the one hand, fleet tracking allows checking if the goods are in the place they have to be at every moment. (see Step 3 in Fig. 2). On the other hand, traceability allows knowing specific details about the path of the goods, detecting possible bottlenecks in order to improve next parts of the route. The on demand request options enable asking the driver's device, which asks to the RFID reader of the container in order to know if everything is fine at that moment. The driver's smartphone returns the answer that can be either OK or Error, and in this case some details about the problem are attached. It is remarkable that only users who are authenticated in the system and with the corresponding permission can ask to the driver's smartphone.

This system also allows to configure goods reception and when the container is next to the reception place, the driver's smartphone will send automatically a message to the responsible of reception. To ensure the privacy and security of data, the server in the cloud is secured. For the implementation of the Web Platform we are using a Node

 ISBN: 978-1-61208-280-6

JS server with Mongo DB for data and express, framework to create the webservices for Node JS. For the front-end we are use Bootstrap framework and GMaps Api.

E. Wi-Fi P2P Request

In order to facilitate authorities work and merchandise management, the app includes an authority interface that agents can use to examine the content of every container without looking inside. (see Step 4 in Fig. 2). In this way, they can check easily if all the information about goods, enterprises and carriers are correct. Otherwise, they do a physical control so that they can check the content and see if it corresponds to the information the system provides. This interface is secured as only authorized people can access to the information through it. In particular, a lightweight P2P authentication method similar to the one presented in [17] is used. This interface will be implemented by using the Wi-Fi Direct APIs for Android.

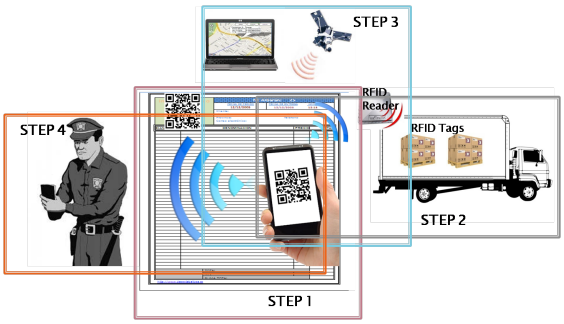


Figure 2. Complete System for Ubiquitous SC.

V. Conclusion and Future Research

In this work, a new logistics system combining different technologies to improve the efficiency and security of the SC is presented. In particular, it allows checking the merchandize not only in the loading and delivery moments, where most problems usually happen, but also at any time, as it offers tracking and tracing of goods. Furthermore, the system's interface to control the merchandise facilitates the authorities' work to perform the merchandise management in custom areas. This tool is its Beta version for the Android platform and will be soon available in the Google play store. Since this is a work in progress, there are many open questions such as the analysis of which are the most appropriate encryption mechanisms for the different communications. The analysis of the time and risks in the proposed SC tool, as well as a comparison between the improvement degree of our proposal and other systems are also future works.

Acknowledgment

Research supported by the MINECO and the FEDER Fund under Projects TIN2011- 25452 and IPT-2012-0585-370000, and the FPI scholarships BES-2009-016774, and ACIISI-BOC Number 60.

References

[1] European Rapex Report, 2011. Available at http://ec.europa.eu/consumers/safety/rapex/ [retrieved: April, 2013]

[2] "C investment hot. U.S. companies have begun to attack the more than $117 billion in excess inventory and $83 billion in lost sales." Material Handling Management, 2004.

[3] D. Giusto, A. Iera, G. Morabito, and L. Atzori, "The Internet of Things," 20th Tyrrhenian Workshop on Digital Communications. Springer, 2010.

[4] A. Brewer, K. J. Button, and D. A. Hensher, "Handbook of logistics and supply-chain management," Emerald, 2001.

[5] M. Christopher, "Logistics And Supply Chain Management: Creating Value-Adding Networks," Financial Times Prentice Hall; Edicin: 3rd Revised edition, 2005.

[6] B. M. Beamon, "Supply Chain Design and Analysis: Models and Methods," International Journal of Production Economics 55(3), 1998 pp. 281294.

[7] TV Rheinland, "Case Study: Supply Chain Security Analysis," Report, 2007.

[8] B. Borgstrm, "Exploring efficiency and effectiveness in the supply chain. A conceptual analysis," The 21st Annual IMP Conference, 2005.

[9] SK. Srivastava, "Green supply chain management: a state-of-the-art literature review," International Journal of Management Reviews 9(1), 2007 pp. 5380.

[10] D. C. Ranasinghe, S. Devadas, P. H. Cole, "A Low Cost Solution to Cloning and Authentication Based on a Lightweight Primitive," Networked RFID Systems and Lightweight Cryptography, Part IV, Springer, 2008, pp. 289-309.

[11] P. Caballero-Gil, C. Caballero-Gil, J. Molina-Gil, C. Hernandez-Goya, "Self-Organized Authentication Architecture for Mobile Ad-hoc Networks", 6th International Symposium on Modeling and Optimization in Mobile, Ad-Hoc, and Wireless Networks, Berlin, Germany, 2008: pp. 01-03

[12] "Package Buddy," Available at play.google.com, [retrieved: April, 2013].

[13] "Package Tracker Pro," Available at play.google.com, [retrieved: April, 2013].

[14] M. Bouet, A. L. dos Santos RFID Tags, "Positioning Principles and Localization Techniques," Wireless Days, 2008. WD '08. 1st IFIP Date of Conference: 24-27 Nov. 2008. pp 1-5.

[15] B. King, X. Zhang, "Securing the Pharmaceutical Supply Chain using RFID," Multimedia and Ubiquitous Engineering, 2007. MUE '07. International Conference. April 2007. pp 26-28.

[16] QR Code Generator, "http://www.qrcode.es/es/generador-qr-code/," [retrieved: April, 2013]

[17] C. Caballero-Gil, P. Caballero-Gil, A. Peinado-Domínguez, and J. Molina-Gil, "Lightweight Authentication for RFID Used in VANETs," EUROCAST 2011, LNCS 6928, pp. 493-500.

 ISBN: 978-1-61208-280-6

Modding and Cloud Gaming: Business Considerations and Technical Aspects

Alexander Wöhrer
St. Pölten University of Applied Sciences
St. Pölten, Austria
email: alexander.woehrer@fhstp.ac.at

Yuriy Kaniovskyi
University of Vienna
Vienna, Austria
email: yk@par.univie.ac.at

Maximilian Kobler
University of Applied Sciences Burgenland
Eisenstadt, Austria
email: maximilian.kobler@fh-burgenland.at

***Abstract*—Cloud computing is changing the IT landscape and enabling new businesses models like Games-as-a-Service (GaaS). The current GaaS approach is characterized by a simple 'shift' of traditional games just provisioned in different ways and neglecting the gamer as an active and participating entity in the overall game development process. This short-paper provides a high level discussion on the recent trends of user-produced game modifications (aka modding) in relation to the GaaS approach. Our contribution is twofold: we first derive the gaming business and industry needs and then, we present the required technical design including the basic security considerations.**

***Keywords*—cloud computing; gaming; modding; business model; security aspects; technical design;**

I. INTRODUCTION

Cloud computing [1], the flexible use of various resources (storage, computation, etc.) delivered as a service over a network on a pay-per-use model, is changing the IT landscape and enabling new business models such as Games-as-a-Service (GaaS) [2]. The general architecture of GaaS followed by OnLive [3] or Gaikai [4], pictured in Fig. 1, shows a strict separation of concerns where the gamer (client side) simply provides the input to the processing service (server side) and later on consumes the subsequent output. The current GaaS approach is characterized by a simple *'shift'* of traditional games provisioned through several different devices, e.g. mobile phone, set-top box or TV. One side-effect of this development is the absence of user-produced game modifications (also known as modding [5]), due to the lock-down of current GaaS offerings. Modding is well studied from the pre-GaaS era. Several types of mods and their in-game support are described in [6], while their impact on the gaming industry is documented in [7]. Nevertheless, a consolidation with and reflection on the recently emerging cloud gaming paradigm is largely missing.

Current research for cloud gaming mainly focuses on topics relating to network issues regarding latency [9]; infrastructure issues concerning data center distribution [10]; dynamic provisioning [11]; and architecture models for 'openness' [8]. However, the common business model of cloud gaming [12] is neglecting the gamer himself as an active and participating entity in the overall game development process. The realization that active gamer participation in the game development process has a great impact factor is out of question. Producers constantly try to predict the addictiveness [13] of games in advance in order to change their future game design accordingly.

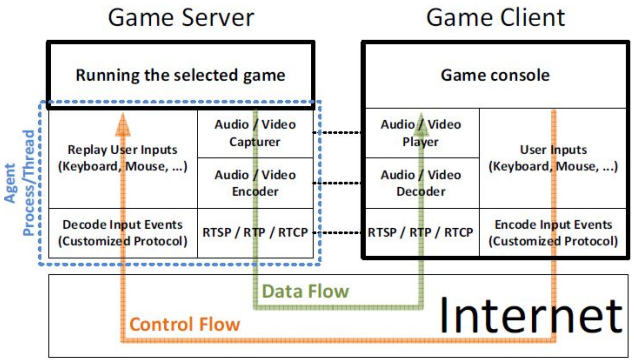


Fig. 1. Modular view of server and client in Cloud gaming [8]

Given the constantly increasing number of competing titles the questions arises: How can game producers attract new gamers, even years after of the title's release, and keep them involved for as long as possible in this new gaming era?

In this paper we propose to embrace the cultural change happening throughout the internet towards participation by supporting modding for Cloud gaming. This notion is especially relevant to titles that promote a creative gaming experience – as games such as Minecraft [14], Spore [15] or LittleBigPlanet [16] not only promote creativity within – but also out of the core game context. Our contribution in this paper towards the support of user-produced game modifications (aka modding) for Cloud gaming is twofold: first, we derive its business need and second, we present the required technical design including security considerations.

The rest of the paper is organized as follows: Section II elaborates on the need to support modding including an extended cloud gaming business model while Section III describes the technical considerations associated with that need. We finish the paper in Section IV with our conclusions and an outline of promising future work in this area.

II. MODDING IN THE CLOUD - THE 'WHY'

Gaming 2.0 [2] – as T. Chang called the recent transition of gaming – describes the major changes of consumer behavior as follows (highlighted here are the most relevant aspects for GaaS):

- rise of the digital natives: users adept at seeking out new offerings and share (one popular form of

 ISBN: 978-1-61208-280-6

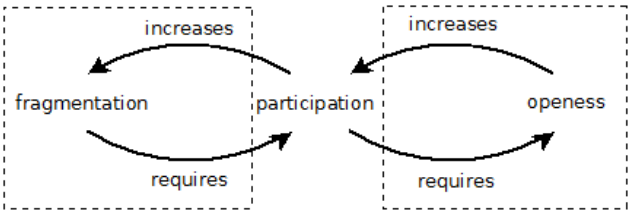


Fig. 2. Interdependencies of Gaming 2.0 factors

participation) their findings.

- irreversible fragmentation and short attention spans: niche offerings replace a few massively popular hits as the web allows every user to find offerings that suit a particular taste, also known as the long tail [17].
- new, open and lightweight platforms: design towards a core game, also known as the meta-game, as a wrapper for mini-games to encourage users to share, level-up, collect, buy/sell/trade/explore and try again.

Let us assess current GaaS offerings against these developments. The required *fragmentation* of game offerings can not be accomplished by long cycle 'big bang' developments by a relatively small group of active game producers compared to a huge group of passive game players. It requires the active *participation* of the gamers in order to attract them in the first place by allowing for user-led innovation [18] or to keep them later on by improving immersion or the overall challenge [19]. The possibilities of the participation itself depends directly on some degree of *openness*. By openness we mean what kind of contributions are made available to the user and supported by the original game developers (e.g. through an API or mod-toolkit). This functionality ranges from simple sound files replacement over the standard ones; visual improvements via new texture files for already defined game objects; cinematic video sequences for storytelling; GUI extensions to completely new levels or mini-games. See last row of Table I for examples. Fig. 2 depicts the relation between these attributes. Let us have a closer look onto the central attribute of the model - namely participation, that is provided by the user-created game content.

Fig. 3 depicts the potentially overlapping groups of involved users in Gaming 2.0 including an indication of their

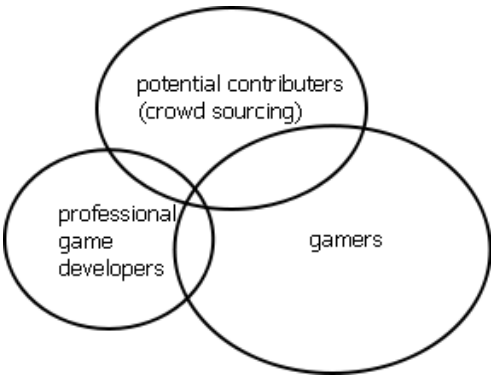


Fig. 3. Potentially overlapping groups of people in Gaming 2.0

dimensions. The largest group, the gamers, will be the consumers of the emerging fragmentation. Contributors, delivering the participation aspect, are driven by the following key motivations [20]: playing (including the identification of improvement of the gaming experience), hacking, researching, artistic expression and co-operation. Co-operation plays an especially important role. Due to the increased complexity of games, the time of the lone-wolf modder is no longer viable. In addition, the online problem-solving and production model, called *crowdsourcing*, is increasingly employed for original thought and increased innovation [21]. Modding differs from traditional crowdsourcing [22] in the outcome belonging to the developing crowd (one to many persons), instead of the software developer who pays the crowd. The smallest group depicted in Fig. 3, the core game developers, are responsible to design and implement the required architecture towards openness. The authors of [23] describe several techniques regarding this particular topic. Additionally, mechanisms for coordinating and facilitating modding teams to exploit the provided openness have to be developed - especially for the more complex mods, shown in Table I. Note the fact that the kernel attribute participation is external to the GaaS providers and game studios, which indicates that the earlier mentioned more complex symbioses [5] between gaming companies and individual (prod)users will have to arise.

III. Modding in the Cloud - the 'how'

Fig. 4 reflects our extended cloud gaming business model. The upper half shows the unchanged GaaS model of a game licensor giving the required binary code to the GaaS provider in order to host the game for remote playing. The end-user (the gamer) chooses its target game title from some portal or menu. The subscription fees are split up as revenue to the GaaS provider and the game licensor. Depicting the business model extension for modding is shown in the the lower half of Fig. 4. The original model is extended by adding a separate revenue cycle. As a prerequisite, some form of development kit or API has to be provided by the game licensor to the GaaS provider. The development kit is hosted by the GaaS provider as well and can be used by end-users, becoming potential contributors. For

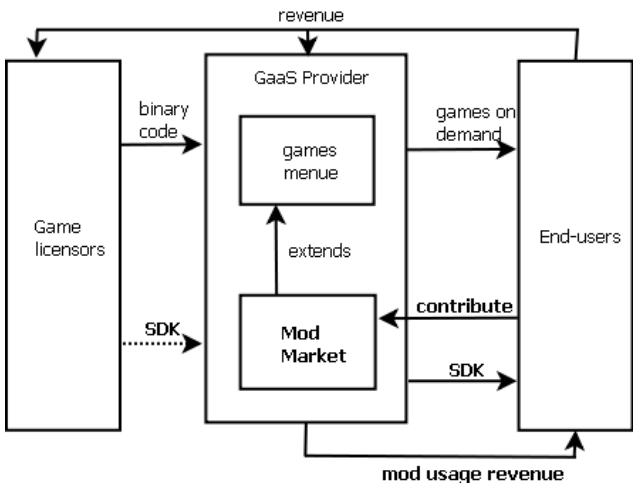


Fig. 4. Extended Cloud gaming business model (adopted from [12]

TABLE I. ASSOCIATED RISK, COMPLEXITY AND EXAMPLE/TOOLS FOR THE DIFFERENT KINDS OF USER-CREATED GAME MODIFICATIONS.

	audio	texture	video	model	GUI extension	level/rules
GaaS risk	low	low	low	low	medium	medium
user risk	low	low	low	low	medium	low
complexity	little	little	medium	medium	much	medium
example/OS tool	own voice/sound recorder	jpeg/gimp	intro/blender	3D item model/blender	auto-aim/-	Counter Strike/-

simplicity, we assume that a contributors use one development account to submit their contribution, the specific user-created game modification, together with a price tag back to the GaaS provider making it available in its Mod-Market catalog. A big difference is the focus on *'usage'* rather than on *'downloads'* as in current mobile app markets like Google Play and Apple AppStore. Popularity of a mod can be expressed much diverse as much more variables are available like numbers of time re-used (or integrated), overall time used, etc.

A gamer can now choose to play a core game extended by one or more mods if he is willing to pay extra for the contribution. Although mods - as such - are usually free, this model considers them to be equivalent to downloadable game content (DLC) - a widely adopted and proven business model in the gaming industry. This additional income is then split up as revenue between the GaaS provider and the mod contributor via a certain ratio, e.g. 70 percent for the mod contributor and 30 percent for the GaaS provider. This ratio is likely to be also dependent on the resource consumption of the mod. While a simple sound or video replacement might not shift it towards the GaaS provider, a resource (e.g. GPU) intensive extension might do so.

One additionaly responsibility of the GaaS provider is to check mods for consistency and compatibility with other mods – similar to CBSE [24] – should the user chose to use several of them together. A mod pre-publishing procedure has to take place, not only to categorize the submission, but also to analyze stability and contradictions in function calls to the core game engine. Some mods, on the other hand, may require additional mods to work. This has to be taken into consideration – in technical and revenue terms – as well.

From a technical perspective the support of modding in the Cloud is associated with three main challenges:

- game instantiation: What changes by introducing mods?
- security aspects: What is the risk of opening up GaaS?
- contribution support and mod deployment: What possibilities and dependencies exist?

Regarding game instantiation, the current high-level approach to game instantiation after a user selected its target title works like a two phase approach: First, find a host machine with the required resources (in terms of storage, computation, graphics capabilities, network parameters). Second, instantiate the (rather static) Virtual Machine (VM) image of that game. Optionally, make the 'saved games' available (copy or link) to the VM instance. Having modding enabled requires a more sophisticated three phase approach. First, calculate the requirements of the VM for the core game and all selected mods. Second, create and instantiate the image. Third, push the configuration changes including the required (additional) digital content onto that VM. The requirements on configuration automation and the network infrastructure are increasing. The latter implies that the larger the required mod data is, the more likely it has to reside at the GaaS provider to allow for efficient game instantiation.

Regarding security aspects – to the best of our knowledge – the current GaaS approach follows a lock-down approach without the possibility to deploy user-created game modifications GaaS-side at all. When considering the security aspects of opening up GaaS, one has to differ between the contribution process and the actual contribution. We will focus on the latter, as we think that mobile-app development has pioneered here already working solutions, e.g. Apple's iPhone app development process. As can be seen in Table I, the risk of *passive* basic building blocks (audio, video, model, texture) for GaaS game titles as well as the user experience is low as the formats are going to be predefined and can be well analyzed. Given a detailed enough target format description and any additional other requirements, e.g. maximum file size, even open source tools could be used for their creation. This changes for the more advanced modifications of *active* parts like GUI extensions and/or new levels as they are typically following an often complex, proprietary internal format and/or programming language, so explicit tools for their creation have to be provided. However, whole eco-systems have already developed around these advanced modifications in the non-GaaS gaming world including premium support offers, e.g. Curse Client for World of Warcraft, Rift, etc.

Regarding contribution support and mod deployment options, the first depends on contribution complexity while the latter depends on mod-size and combined availability. Making significant user-produced contributions typically requires multiple people joining forces and the availability of complex productions-lines and associated tools [23] for integrating and testing them. Recent browser-based IDE-as-a-Service [25] offerings provide installation-free usage and tight collaboration support on a project. This leaves the decision where the contribution support shall reside. Resource management and availability are both core competences of GaaS providers while the desire for GaaS provider independence favours a deployment option under the control of the game licensor. The decision for a mod deployment option, e.g. all at GaaS provider versus distributed over multiple systems, can be reduced to a dependence on mod-size and system availability. Assuming large mod sizes makes it unpracticle to transfer them to the GaaS provider each time it is requested. Additionally, if failure of a system part leads to the combination (core game A_x + mods A_y) becoming inoperable, the two parts are considered to be operating in series leading to availability $A = A_x * A_y$. From that well known formula follows that the combined availability of two components in series is always lower than the availability of its individual components. Assumming that the core game A_x at the GaaS provider has a very high

 ISBN: 978-1-61208-280-6

availability as it is its core competence strongly argues against distributing mod data on multiple systems.

IV. Conclusion

This paper argues that gaming trends tend to drift more and more towards user participation for creating new contents within games, which implies the need to adapt cloud gaming to open up and to support the required fragmentation. We derived the business need to support modding in the Cloud, identified the challenges and interdependencies of the Gaming 2.0 factors and developed an extended business model for Cloud gaming that includes user participation through modding. Additionally, we described the technical challenges associated with it depending on mod-size and contribution complexity as well as security considerations depending on mod-type.

Promising future work and opportunities in this field include the analysis of the formal business model, in particular regarding the fine grained pay-per-use concepts of the Cloud on real usage/appearance rather than on simple inclusion; and the elaboration of a description model for hierarchically structured mods, where one is dependent on another.

Acknowledgment

The authors would like to thank Martin Lukic and Martin Ivancsits, both students at University of Applied Sciences Burgenland, for their input and valuable comments.

References

[1] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the clouds: A berkeley view of cloud computing," University of California at Berkeley, Tech. Rep., 2009.

[2] T. Chang, "Gaming will save us all," *Commun. ACM*, vol. 53, no. 3, pp. 22–24, Mar. 2010.

[3] OnLive, http://www.onlive.com/.

[4] Gaikai, http://www.gaikai.com/.

[5] O. Sotamaa, "The player's game: Towards understanding player production among computer game cultures," PhD Thesis, University of Tampere, Finland, 2009.

[6] W. Scacchi, "Modding as a basis for developing game systems," in *Proceedings of the 1st International Workshop on Games and Software Engineering*. ACM, 2011, pp. 5–8.

[7] C. Camargo, "Modding: changing the game, changing the industry," *Crossroads*, vol. 15, no. 3, pp. 18–19, Mar. 2009.

[8] C.-Y. Huang, C.-H. Hsu, Y.-C. Chang, and K.-T. Chen, "GamingAnywhere: An open cloud gaming system," in *Proceedings of ACM SIGMM Conference on Multimedia Systems (MMSys'13)*, 2013.

[9] S. Choy, B. Wong, G. Simon, and C. Rosenberg, "The brewing storm in cloud gaming: A measurement study on cloud to end-user latency," in *Network and Systems Support for Games (NetGames), 2012 11th Annual Workshop on*, nov. 2012, pp. 1 –6.

[10] S. Choy, G. Simon, B. Wong, and C. Rosenberg, "Edgecloud: A new hybrid platform for on-demand gaming," University of Waterloo, Tech. Rep. CS-2012-19, 2012.

[11] M. Marzolla, S. Ferretti, and G. D'Angelo, "Dynamic resource provisioning for cloud-based gaming infrastructures," *Comput. Entertain.*, vol. 10, no. 3, pp. 4:1–4:20, Dec. 2012.

[12] A. Ojala and P. Tyrvainen, "Developing cloud business models: A case study on cloud gaming," *Software, IEEE*, vol. 28, no. 4, pp. 42–47, 2011.

[13] J.-K. Lou, K.-T. Chen, H.-J. Hsu, and C.-L. Lei, "Forecasting online game addictiveness," in *NetGames*, 2012.

[14] Mojang, "Minecraft," https://minecraft.net accessed 03/04/2013, 2009.

[15] Maxis, "Spore," http://www.spore.com accessed 03/04/2013, 2008.

[16] Media Molecule, "LittleBigPlanet," http://littlebigplanet.com accessed 03/04/2013, 2008.

[17] C. Anderson, *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion, 2006.

[18] National Endowment for Science, Technology and the Arts, "The new inventors: how users are changing the rules of innovation," pp. 1–48, 2008.

[19] J. Milton, "A comparison and analysis of techniques used in computer games and interactive fictions aimed at engaging users over a period of time," in *Interactive Multimedia Conference*, 2013.

[20] O. Sotamaa, "When the game is not enough: Motivations and practices among computer game modding culture," *Games and Culture*, 2010.

[21] A. Kittur, "Crowdsourcing, collaboration and creativity," *XRDS*, vol. 17, no. 2, pp. 22–26, Dec. 2010.

[22] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce*, ser. EC '10. ACM, 2010, pp. 209–218.

[23] D. S. Batory, C. Johnson, B. MacDonald, and D. von Heeder, "Achieving extensibility through product-lines and domain-specific languages: a case study," *ACM Trans. Softw. Eng. Methodol.*, vol. 11, no. 2, pp. 191–214, 2002.

[24] D. C. Craig, "Compatibility of software components: Modelling and verification," PhD Thesis, Memorial University of Newfoundland, Canada, 2007.

[25] T. Aho, A. Ashraf, M. Englund, J. Katajamki, J. Koskinen, J. Lautamki, A. Nieminen, I. Porres, and I. Turunen, "Designing IDE as a service," *Communications of the Cloud Software*, vol. 1, no. 1, 2011.

 ISBN: 978-1-61208-280-6

Three-Tiered Data Mining for Big Data Patterns of Wireless Sensor Networks in Medical and Healthcare Domains

Jong P. Yoon

MATH/CS Dept, Cybersecurity Program
Mercy College, Dobbs Ferry, NY 10522, USA
jyoon@mercy.edu

***Abstract*— As smartphones become an emerging interface platform between humans and systems, they also enable wireless sensors to interface with host servers. As sensors monitor application domains and sensor data is frequently polled and transmitted to a host server, the server data will be a big volume, big variety and big velocity, which is the characteristic of big data. Mining patterns from big data is a very important and active research topic since it can be used to forecast and "nowcast" for any dynamisms in application domains. However, typical data mining algorithms are not successful yet due to the characteristics of big data. This paper describes three-tied data mining paradigm. Alongside the streamline of sensor data transmission, at the microcontroller tier, sensor data sets are mined to form patterns, at the smartphone tier, negative and positive patterns are grouped and verified, and finally at the host server tier, human expertise is associated with the patterns. The contribution includes 1) lowering data transmission by mining from the lower tiers, 2) mining time-critical data earlier than it would be done at the host server tier and 3) hence urgent responses can be made timely at the proper tier.**

Keywords- data mining; microcontroller;smartphone; wireless sensor networks; big data

I. INTRODUCTION

Big data is a large collection of complex data sets that are dynamically generated from various sources. It is therefore inefficient in processing or managing using traditional technologies. Data becomes large in volume if it is generated from various sensor networks. It should be transmitted in high speed if such data sets are processed centrally in a host. Wireless sensors monitor and poll various sensor data, which is then transmitted to a host server for further analysis and management. The data collected in a host is not only stored safely for archiving but it will also be analyzed to generate abstracted patterns or to associate meaning patterns for decision assistance.

In wireless sensor networks (WSNs), sensor nodes are deployed in unattended or less-attendable (nurses do attend and watch over patients 24/7) environments where wireless communication becomes more effective. Less-attendable hospital patient room environments may produce large quantities of data, which will be more dynamic if it is polled from various sensor nodes. Wireless patient monitoring devices are equipped with sensors capable of monitoring specific vital sign datasets [1,2,3]. For example in Figure 1, Electroencephalography (EEG) sensors are monitoring patient's brain activities, Electrocardiography (ECG) sensors are monitoring the function of patients' heart, a blood oxygenation sensor is equipped with IV (Intraventricular) for a patient while a motion detector monitors the patient and any infrared (IR) or ultrasonic sensors. Wireless sensors monitor patients, traditional medical equipment tools, and room environments [4]. Such wireless sensors as temperatures, pressure, moisture, chemical updates, images and audios, etc, can be deployed very easily and cost-effectively in medical treatment and healthcare, as they are running on microcontrollers such as Arduino [5] or its clones.

The radio signals polled by sensors are formed to be a sensor data. Sensor data polled from WSNs is then transmitted to the relevant medical practitioner and/or a host machine (see Table 1). A host machine may be a server or a cloud computing environment. Moreover it can often be a smartphone, which is widely and ubiquitously used especially in the medical and health communities. Sensor data or the outcome of sensor data processing will drive staff and doctors towards greater efficiency and quality of medical and health treatment. Healthcare and medical data will constitute big data [6,7], which will be more significantly big data if it is collected from WSNs.

One of the promising methods for sensor data analysis is data mining (DM). However, typical DM methods or DM technologies on integrated datasets [8] are not satisfactorily applied to the big data of sensor data sets. It is in part because sensor data of medical and health WSNs is

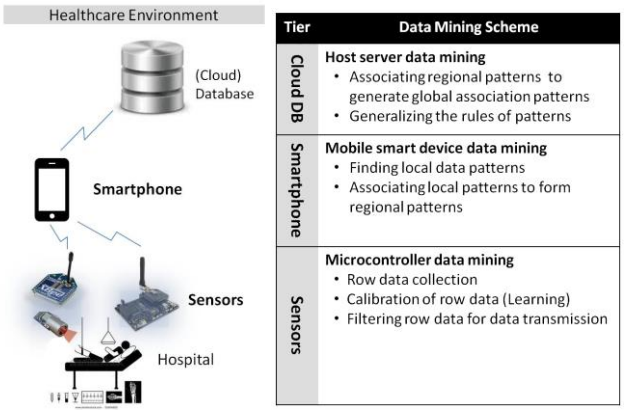


Figure 1. Sensor data streaming in hospital example, and data mining techniques along the sensor data streaming

 ISBN: 978-1-61208-280-6

streaming in dynamically from various sensor sources. The integrated volume of big data and the high-speed velocity of streaming WSNs make it inefficient for the DM to produce patterns that specific enough to correspond to the patients' responses in real time. If entire sensor data sets are integrated centrally in a host machine, sensor-sensitive local states are not efficiently taken into consideration and local exigent responses become impossible. In general, big data becomes difficult to process using on-hand database management tools or traditional data analytic applications. Simply speaking, automatically generated big data is almost impossible to be analyzed in servers' main memories only.

As such, this paper proposes a new wireless sensor healthcare approach to combine two architectural streams together between WSNs and DMs as shown in Figure 1. In WSNs on the left side, 1) multiple wireless sensor data is polled by microcontrollers, which will then transmit to 2) mobile smart devices and further to 3) host machines. In DM on the right side of the figure, multilevel DMs are proposed.

The contribution of the proposed research approach includes 1) lowering data volume transmitted from WSN node sensors and the host server where data is processed, 2) mining and analyzing sensor data early, in that the node specific actuation can be taken quick into consideration.

The remainder of this paper is organized as follows. Section 2 describes the sensors available for monitoring patients in healthcare and medical domains and some issues that we have encountered using them. Section 3 describes the three-tiered DM processes: 1) microcontroller DM at the sensors level, 2) mobile smart device DM at mobile phones level, and 3) server DM at a host or a cloud server level. Section 4 describes implementation details. Finally, Section 5 describes the conclusion and future work.

II. PRELIMINARIES

This section introduces sensors that can monitor patients for healthcare, microcontrollers that can operate, manage and wirelessly transmit sensor data, and smartphone devices.

A. Sensors

Many of sensors used in healthcare and medical services are traditional medical sensors: EEG, EMG, ECG, PPG, SpO2, etc. EEG (Electroencephalography) is a test that measures and records the electrical activity of brains. Neural oscillations are observed in EEG activities. EMG (Electomyography) measures the electrical activity produced by skeletal muscles, using a motor unit. When a motor unit fires, the impulse (i.e., action potential) is carried down the motor neuron to the muscle. ECG (a.k.a EKG, Electrocardiography) measures the electrical activity of the heart over a period of time. There are 12-lead ECG electrodes plus more to improve the sensitivity in detecting myocardial infarction involving territories not normally seen well. Blood pressure or more specifically Wearable blood pressure sensor [9] is a device that can monitor and measure and it could help diagnose hypertension and heart disease. This device uses pulse wave velocity, which allows blood pressure to be calculated by measuring the pulse at two points along an artery, and it monitor the hydrostatic pressure changes to prevent high blood pressure, which is a common risk factor for heart attacks. PPG (Photoplethysmography) measures pulse oximeters by cardiovasculary monitoring. SpO2 (Oxygen Saturation) measures the percentage of hemoglobin binding sites in the bloodstream occupied by oxygen.

Each such medical and healthcare monitoring equipment is equipped with the basic IR sensors. IR sensors include chemical sensors, acoustic and sound sensors, electric current and magnitude sensors, weather and moisture sensors, ionizing radiation and subatomic particles sensors, pressure sensors, distance and position sensors, force and level sensor, etc. Such IR sensors are loaded on microcontrollers and so they can be a wireless sensor as well as programmable.

B. Microcontrollers

A microcontroller is a single integrated circuit, which consists of a CPU, memory and programmable IO peripherals. It allows the firmware to handle interrupts in response of the events of sensors. There is a dedicated pulse width modulation (PWM) block, which makes it possible for the CPU to control power converters, resistive loads, motors, etc, without using lots of CPU resource. Sensor data can be periodically polled, but not stored in a microcontroller, since most microcontrollers have a limited storage (memory) space. Serial peripheral interface (SPI) and universal asynchronous receiver/transmitter (UART) can receive and transmit sensor data from/to external devices such smartphones.

For example, Arduino [5] consists of a simple open source hardware board with Atmel ARM or Atmel AVR, and a programming language (e.g., C and C++) compiler with a boot loader. A microcontroller board can support the aforementioned sensors as an Arduino shield. Example of communication shields includes wifi (IEEE 802.11) shields, Bluetooth (IEEE 802.15.1) shields, XBee shields (IEEE 802.15.4) [10], GPS shields, etc.

C. Smartphones

A mobile phone has also various sensors and communication components running on a mobile operating system such as Google's Android or Apple's iOS. The sensors equipped with most smartphones are proximity sensor, GPS sensor, cameras, accelerometer, etc. There are network protocols available, e.g., 3G/4G wireless communication, IEEE 802.11 wifi and 802.15.1 bluetooth, etc. As described in the previous subsection, since microcontrollers have insufficient storage spaces and they have the same network protocols, through IEEE 802.11 or 802.15 sensor data can be transmitted to smartphones.

 ISBN: 978-1-61208-280-6

As such, the sensor data received from microcontrollers is stored in smartphones. Note that smartphones have an embedded database, more precisely a database library called SQLite. A smartphone has a storage capacity that is large enough to store sensor datasets in SQLite.

A rule of thumb is to assign one microcontroller per a patient or a patient room, where multiple wireless sensors deployed to multiple spots of patient body. Each medical or healthcare staff holds a mobile smart device, which polls sensor data from microcontrollers as they approach the microcontrollers.

III. THREE-TIERED DATA MINING

Since typical DM techniques are not satisfactorily applied to monitoring, mining and analysis in wireless healthcare sensor networks, this section describes three-tiered data mining. As shown in Figure 1, three DM models are available from the patient body, to the regional tier like each hospital floor, and the global tier like an entire hospital. The patterns mined from multiple microcontrollers at a patient body-tier are associated together on a smartphone at a regional tier, which will then be more generalized to form a discovery rules at the global tier server.

Note that this paper does not propose a new algorithm of data mining, but proposes a new paradigm of data mining in big data that is collected along the streamline of data transmission from wireless sensors to wireless smart devices and to host machines. Before proposing the three tiered-data mining paradigm, it is assumed that all sensor datasets are cleansed and trusted.

A. Microcontroller Data Mining

As described in Section 2, one or more sensors are plugged in on a microcontroller, which is called a WSN node. Multiple sensor data can be collected by a microcontroller at each every single polling [12]. For example in Figure 1, a WSN node may control EEG, ECG, moisture and pressure sensors. EEG and ECG respectively monitor the brain and heart activities of a patient, while the moisture sensor monitors the IV injection and the pressure sensor monitors the patient's bed. In this very common situation, two phases of data mining are proposed: Training and Calibration Phase and Pattern Transmission Phase.

1) Training and Calibration Phase: Each WSN node needs to be calibrated and the correlation of sensors needs to be identified by medical experts. Since it is unnecessary to poll data from all sensors at the same time in the same interval, a microcontroller should set a polling time for each sensor.

Suppose that four sensors, *s*1, *s*2, *s*3, *s*4 are deployed over a human body. Depending on the disease of patients, a different sensor set will be deployed. The sequence of such sensors is trained by calibration or determined by medical experts. A sequence of sensor data is written in regular expressions. A polling sequence (*s*1, (*s*2, *s*3)*, *s*4)* collects data from *s*1, and then *s*2 and *s*3 for multiple times, finally *s*4. This sequence is iterated.

For example, consider the following Arduino C programming of the nested iterations where four sensor datasets are polled in different frequency:

```
void loop() {
  sensorVal1 = analogRead(flexiForce);
  for (int i=1; i<=2; i++) {
    sensorVal2 = digitalRead(EEG);
    delay(100);
    sensorVal3 = digitalRead(ECG);
  }
  sensorVal4 = analogRead(liquidFlow);
}
```
(1)

The polling sequence of Expression (1) is (flexi force sensor, (EEG, ECG)*, liquid flow sensor)*, and more precisely, the iteration of the inner loop is 2. It means that between the flexi force and the liquid flow sensors are polled, a sequence of data polling from EEG and ECG occur twice.

2) Pattern Transmission Phase: From a sequence of polling sensor data discussed in the previous subsection, the pattern can be very easilly formed in the same expression of the polling sequence. The pattern obtained from the polling sequence is very similar to the pattern associated by A Priori association algorithms [11]. In the algorithm, an association pattern can obtained if the pattern has enough supports or frequency. In the same spirit, revisiting the sequence polling once more here, the pattern (s1 & s4) is formed since there exists a supportive pattern, (s2 & s3) with enough frequency. An example of pattern from Expression (1) is as follows:

```
(flex force = 0.1 mV) &
(liquid flow = 1000 nanoliter/min)

With the evidence of
(value of EEG = 60 microvolt/Hz) &
(value of ECG = 160 heartbeats)
```
(2)

Above Expression (2) is about a situation such that the flex force of patient's bed is almost negligible and the liquid flow rate of patient's IV is far greater than a nano liter per minute. This is indicative that a patient fell down from the bed and the IV injection is disconnected from the patient. This situation happens in many hospitals: at night a patient has abnormal pains, which is detected by EEG and ECG, and with no observation of medical staff, he or she moves from the bed and falls down. This risk of falling is a serious patient safety issue and a timely responsive actuator to a patient fall could help to minimize the negative repercussions of the fall.

The primary point of this research is that wireless sensor networks should be established and properly maintained and managed.

The beauty of this approach is that the pattern can be transmitted to a smartphone without waiting for more events

 ISBN: 978-1-61208-280-6

to count the enough frequency of *s*1 and *s*4. Only with the frequency of *s*2 and *s*3, e.g., in this case, after two occurences only, the pattern is quickly uploaded to a smartphone. If it is an urgent case, on the smartphone some additional alerts will be made, e.g., calling for a medical specialty. In a more traditional approach to patient care, the alerting of a patient urgently in need of attention may not be as timely.

B. Mobile Smart Device Data Mining

One example of widely used mobile smart devices is smartphones. Smartphones can communicate with wireless sensor networks. They receive sensor data from and/or transmit data or software to WSNs. Note that software packages can be transmitted to WSNs for several reasons [13], one of which is to upgrade. The goal of DM in mobile smart devices is to find correlations among wireless sensors, some from human bodies and others from hospital equipment and environments.

A smartphone can receive patterns, e.g., Expression (2), from one or more microcontrollers, each of which polls sensor data from one or more wireless sensors. Since each patient has different diseases and different symptoms, sensor data polled and correspondingly the patterns collected will be different. As such, at any point in time, it is likely there may be some patterns that are opposite of or conflicting with another. It may be in the case that (*s*1 & *s*4 & *s*5) and (*s*1 & *s*4 & ¬*s*5), where *s*# notes a sensor pattern and the symbol ¬ is a negation. If those two patterns are in the same database, due to conflict, no further reasoning is possible. However, they can be in two different databases for the DM purpose since they are about two different diseases, both will be an important mining factor. Moreover, for verification purposes in DMs, both conflict patterns can be considered together.

Suppose that there are patterns collected from microcontrollers: (*s*1 & *s*4 & *s*5), (*s*1 & *s*3 & ¬*s*5), (*s*4 & *s*5), (*s*3 & ¬*s*5), (*s*2 & *s*4 & *s*5), (*s*3 & ¬*s*5). We can split them into two groups: *positive example*, i.e., one with *s*5 and *negative example*, i.e., another with ¬*s*5. Thus, {*s*1 & *s*4, *s*4, *s*2 & *s*4} for *s*5 and {*s*1 & *s*3, *s*3, *s*3} for ¬*s*5 are obtained. The maximum common factor for each pattern set will be {*s*4} for *s*5 and {*s*3} for ¬*s*5.

However, if (*s*3 & *s*4 & ¬*s*5) is also included in the above scenario, the outcome will change. Since two pattern sets are {*s*1 & *s*4, *s*4, *s*2 & *s*4} for *s*5 and {*s*1 & *s*3, *s*3, *s*3, *s*3 & *s*4} for ¬*s*5, {*s*3} for ¬*s*5 is sound, but {*s*4} for *s*5 is not in the logic. The latter is untrue in the logic but may possibly be true in medicine; it should be notified to medical staff. The logical verification is not always true, but depends on real-world application domains. Therefore, the following definition is used in general for mobile smart device DM.

Definition 1 (*Positive and Negative Examples*) If there exists a pattern and its negation, called the pivot pattern, a set of patterns can be split into two groups. Positive example is a set of patterns that with the positive pivot pattern, and negative example is a set of patterns with negative pivot pattern. DM is performed in the positive example, while verification of DM in the negative example. (3)

Example 1: Consider the following patterns that are collected from microcontrollers.

```
(flex force = 0.1 mV) &
(liquid flow = 1000 nanoliter/min) &
(respiratory rate = 75 bpm)

(liquid flow = 1100 nanoliter/min) &
(respiratory rate = 72 bpm)

(heart beat rate = 160) &
(respiratory rate = 40 bpm)
```
(4)

It is known that normal healthy people have a respiratory rate between 10 and 45 breaths per minute. Hence, Expression (4) can be rewritten as follows:

```
(flex force = 0.1 mV) &
(liquid flow = 1000 nanoliter/min) &
¬(respiratory rate = OK)

(liquid flow = 1100 nanoliter/min) &
¬(respiratory rate = OK)

(heart beat rate = 160) &
(respiratory rate = OK)
```
(5)

According to Definition 1, the pivot pattern is `(respiratory rate = OK)` based on which Expression (5) can be split into positive example and negative example as follows:

```
(heart beat rate = 160) &
(respiratory rate = OK)
```
(*positive example*)

```
(flex force = 0.1 mV) &
(liquid flow = 1000 nanoliter/min) &
¬(respiratory rate = OK)

(liquid flow = 1100 nanoliter/min) &
¬(respiratory rate = OK)
```
(*negative example*)

The minimum pattern in the positive example is `(heart beat rate = 160)`, and the one in the negative example is `(liquid flow >= 1000 nanoliter/min) & (liquid flow <= 1100 nanoliter/min)`. Since any of these patterns can appear in the opposite example, the verification is done, and therefore

`(heart beat rate = 160)` for `(respiratory rate = OK)` (6)

`(liquid flow >= 1000` for `¬(respiratory` (7)

```
nanoliter/min)        &        rate = OK)
(liquid flow <= 1100
nanoliter/min)
```

Above Expressions (6) and (7) are obtained and sent (or texted) to appropriate medical and healthcare staff. If the mined patterns are meaningful, the staff may be able to take an action of treatment.

C. Host Server Data Mining

A host server is an enterprise system that holds not only wireless sensor-related data and patterns but also historical and statistical databases about hospitals.

The patterns formed in smartphones and taken care of by medical staff are uploaded to a server with their consequences. With patterns and consequences, a server can mine more useful and complete rules that may improve the overall healthcare program.

For example, back to the previous example as shown in Expressions (6) and (7), the following rule can be obtained:

```
(liquid flow >= 1000 nanoliter/min) &
(liquid flow <= 1100 nanoliter/min) &
¬(respiratory rate = OK) → emergency
treatment (E1, E2, E3)                    (8)

treatment (E1, E2, E3) → fail

treatment (E1, E3) → success
```

If a patient falls down from a bed, then a sequence of three treatments E1, E2 and E3 needs to be taken. Note that the treatment E# can be associated from the medical historical and statistical databases, which is omitted in this paper.

IV. IMPLEMENTATION ISSUES

This section describes some implementation issues along the streamline of sensor data transmission. Figure 2 shows three possible communication protocols: IEEE 802.11, 802.15.1 and 802.15.4. Smartphone Apps possessed by a medical staff initiate to find and connect Bluetooth devices, which run on microcontrollers (denoted as ① in Figure 2). One microcontroller can communicate with another (denoted as ② in Figure 2) or to the Smartphone Apps (denoted as ③ in Figure 2). Smartphone Apps then communicate with host servers (denoted as ④ in Figure 2).

The following code segments illustrate how an Android phone opens a communication session and communicates with Arduino BTshild. The key segments include the method findBT() and openBT(), then sending and receiving data.

```
void findBT() {
  mBTA = BluetoothAdapter.getDefaultAdapter();
  if(mBTA == null) {
     myLabel.setText("No BTA Available"); }
  if(!mBTA.isEnabled()) {
     Intent enableBT = new
Intent(BluetoothAdapter.ACTION_REQUEST_ENABLE);
      startActivityForResult(enableBT, 0);
  }
  Set<BluetoothDevice> pairedDevices =
mBTA.getBondedDevices();
  if(pairedDevices.size() > 0) {
   for(BluetoothDevice device : pairedDevices){
    if(device.getName().equals("myBT")){
       mmDevice = device;
       break; }
   }
  }
  myLabel.setText("BT Found …");
 }

void openBT() throws IOException {
  UUID uuid = UUID.fromString("00001101-0000-
1000-8000-00805f9b34fb"); // standard
  mmSocket =
mmDevice.createRfcommSocketToServiceRecord(uuid);
  mmSocket.connect();
  mmOutputStream = mmSocket.getOutputStream();
  mmInputStream = mmSocket.getInputStream();

  beginReceivingTransmittingData();
    myLabel.setText("Bluetooth Open…");
}
```

In response to the Android App's request, the hospital room's Arduino microcontroller takes actions: polling sensor data and transmitting the data to Android, as shown in the following code segment.

```
void loop() {
    while (bluetooth.available() == 0);
    fromPhone = (int)bluetooth.read();
    if (fromPhone >=10) {
      analogRead(sensor1);
      delay(1000);
    } else {
      analogRead(sensor2);
    }
   toXB2 = us.Ranging(CM);
    jServo.write(i);
    delay(100);
  }
  Serial.print(toXB2);
  if (Serial.available() > 1) {
    fromXB2 = (int)Serial.read();
    if(fromXB2>=10)
           digitalRead(sensor3);
    else digitalRead(sensor4);
    delay(5000);
  }
}
```

Note that the Arduino code also communicates with an XBee device (denoted as ② in Figure 2), which may sit on the same microcontroller or another separately.

The communication and data transmission paradigm illustrated in Figure 2 can be applied to various application domains. This section deploys 10 mobile WSN nodes in a closed space and illustrates how they identify the layout of the space. Each WSN node consists of an Arduino microcontroller, XBee transceiver antenna, a proximity sensor built on a servo motor and a time synchronizer as

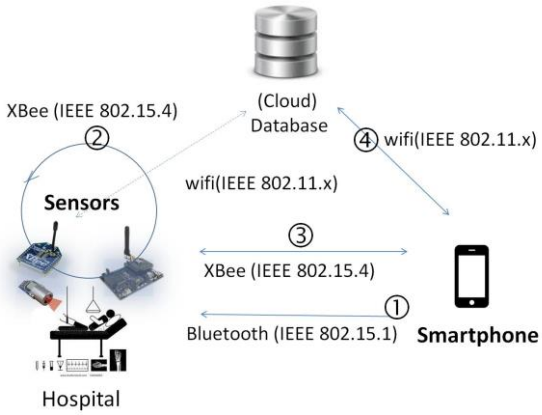
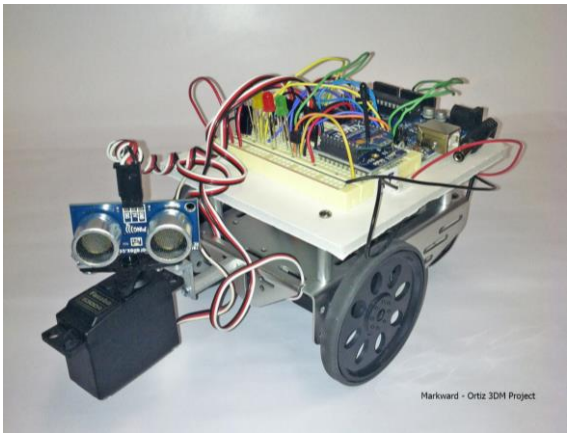


Figure 2. Communication among elements

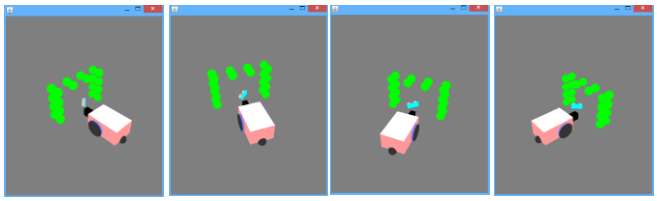
shown in Figure 3 (a). Each WSN node polls a sensor data about the objects that it may identify. Sensor data is transmitted to smartphones in the second tier, and finally reached at a host server in the third tier. Each WSN node and objects that are identified by the node are visualized at the server tier.

The sensor data polled from each WSN node is transmitted and finally reached at the server. Figure 3(b) illustrates the visualized views of the analyzed sensor data transmitted from four WSN nodes. The sensor data can be structured in 7 elements, (*id*, *x*, *y*, *z*, *a*, *d*, *t*), where *id* is the ID of a WSN node, *x*, *y* and *z* denote latitude, longitude and altitude, *a* is an angle, and *t* is the time of polling. The elements, *x*, *y* and *z* can be either an absolute geocode if GPS sensor is used, or otherwise relative coordinators. Given such sensor data, an object is recognized in the server. The data structure of such an object is structured in 4 elements, (*x*, *y*, *z*, *t*), where *x*, *y* and *z* are geocode as above and *t* is the time. The time *t* can be either synchronized if a time sync device is installed. Unless otherwise, it will be a local time. A typical approach is to transmit to the server all the sensor data polled. Note that in our experiment, each node polls and ships out the data at every 1000 msec. The smaller the time interval is set, the more the sensor data can be polled. The experiment presented in this paper does not use GPS devices and the system clocks in microcontrollers are not synchronized.

As discussed, filtering conditions are defined at each WSN node (microcontroller) or a smartphone. The filtering condition defined in each node filters the sensor data. The filter condition used in this experiment is: "if there are two data records polled by the same sensor point out the same object, only one data record is transmitted to smartphones." Another type of filtering condition is defined in smartphones. An example of such conditions is: "if there exist two data records transmitted from two different WSN nodes point to the same object, only one data record is transmitted to the server." This paper shows the experiment with the filtering conditions in WSN nodes only.



(a) WSN node – real view



(b) Visualized views of the objects identified by each WSN node

Figure 3. An implementation of WSN

Another factor for monitoring and sensing objects is the specification of sensors. In our experiment, the factor is the range of proximity sensors. There are three ranges of distance measurement such as long-, medium- and short-distance measuring ranges. The long-distance measuring range of proximity sensors can cover beyond 1m, while the medium and short cover up to 100cm and 10cm, respectively. Sharp proximity of the short-distance measuring range sensor has been used in our experiment.

With the sensor conditions and the system configuration stated above, the preliminary outcome is shown in Table 1. It shows that the data transmission is reduced by 70% by the simple filtering condition at WSN node. Since there are numerously many sensors to be deployed in a real-world application (in military or in health care), more sensor data would be polled. Depending on the bandwidth of wireless networks, it may not be possible to transmit all the data polled. As such, the technique proposed in this paper makes it possible to reduce the sensor data transmission substantially and also to save the power required for WSN nodes.

Moreover, in an experiment with 10 WSN nodes, Table 1 shows that not all sensor data can be received at a higher tier. It is partly due to overwhelmingly big size of sensor data. Even in this case, however, the very same technique proposed here enables all essential data to reach at a higher tier in the proposed technique. Note that the table shows two cells with no numbers but N/A, meaning that not all the data records are reached at the server. In addition to the reduction of wireless sensor data reduction, our experiment shows that

 ISBN: 978-1-61208-280-6

Table 1. Data Transmission – Preliminary Result

	Data received in 1 min		Data received in 10 min	
	# of Records	Size (KB)	# of Records	Size (KB)
Sensor data from a WSN node without Filtering	34	278	345	2987
Sensor data from a WSN node with Filtering	8	64	57	419
Sensor data from 10 WSN nodes without Filtering	351	N/A	3891	N/A
Sensor data from 10 WSN nodes with Filtering	74	577	721	5548

there is no significant difference in patterns mined from between the sensor data with and the one without.

V. CONCLUSION

This paper described a three-tiered data mining from big data sets of wireless sensors deployed to medical and healthcare domains. From wireless sensors over hospital patients to a host server, three major agents are described: microcontrollers, smartphones and host machines. At each such device, data mining paradigm is investigated. Not only reducing sensor data by filtering, but also mining wireless data patterns is performed at the sensor data sets at a lower tier, i.e., at each WSN node.

The contribution of the proposed research approach is a transmission data reduction and timely data mining: 1) enabling WSNs to filter sensor data at the microcontroller tier and 2) reducing data transmission by transmitting time-critical data only to the smartphone, which then aggregate sensor data from various WSN nodes to conduct more efficient data mining.

ACKNOWLEDGMENT

This research is partially funded by Mercy College faculty development program. The author would like to thank James Markward and Juan Ortiz for the preliminary sensor robot implementation and Dr. Christopher Frenz and Dr. Petre Dini for their comments.

REFERENCES

[1] R. Naima and J. Canny, The Berkeley Tricorder: wireless health monitoring, Wireless Health '10, October 2010, pp. 212-213.

[2] J. Woodbridge, A. Nahapetian, H. Noshad, M. Sarrafzadeh, and W. Kaiser, Wireless Health and the Smart Phone Conundrum, SIGBED Review 6(2), 11, 2009.

[3] A. Mainwaring, J. Polastre, R. Szewczyk, D. Culler, and J. Anderson, Wireless Sensor Networks for Habitat Monitoring, ACM Int'l, Workshop on Wireless Sensor Networks and Applications, Septermber 2002.

[4] MIT Wireless Center, http://wireless.csail.mit.edu/index, Successful human tests for first wirelessly controlled drug-delivery chip, http://web.mit.edu/newsoffice/2012/wireless-drug-delivery-0216.html; MIT's CSAIL launches new center to tackle the future of wireless and mobile technologies, http://web.mit.edu/newsoffice/2012/wireless-research-center-founded-1011.html retrieved in Dec., 2012.

[5] Arduino, http://www.arduino.cc/ retrieved in Aug., 2012.

[6] J. Manyika, et al, Big data: The next frontier for innovation, competition, and productivity, McKinsey Global Institute, http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, retrieved in May 2011.

[7] J. O'Donoghue and J. Herbert, Data management within mHealth environments: Patient sensors, mobile devices, and databases, ACM Journal of Data and Information Quality, vol. 4, no. 1, Octover 2012, pp. 1-20.

[8] J. Lockhart, G. Weiss, J. Xue, S. Gallagher, A. Grosner, and T. Pulickal, Design considerations for the WISDM smart phone-based sensor mining architecture, SensorKDD'11, August 21, 2011, pp. 25-33.

[9] MIT, Wearable blood pressure sensor offers 24/7 continuous monitoring, http://web.mit.edu/newsoffice/2009/blood-pressure-tt0408.html, 2009.

[10] XBee, http://www.digi.com/xbee/ retrieved in January 2013.

[11] R. Agrawal and R. Srikant, Fast algorithms for mining association rules in large databases, Int'l Conf. on VLDB, 1994, pp. 487-499.

[12] V. Chandola, O. Omitamou, A. Ganguly, R. Vatsavai, N. Chawla, and J. Gaber, Knowledge discovery from sensor data, SIGKDD Explorations Newsletter, 12(2), 2011.

[13] S. Zhu, S. Setia, S. Jajodia, and P. Ning, Interleaved hop-by-hop authentication against false data injection attacks in sensor networks, Transactions on Sensor Network, 3(3), 2007.

 ISBN: 978-1-61208-280-6

A Novel Risk-based Approach for Online Community Management

Bassem Nasser, Vegard Engen, Simon Crowle, Paul Walland
IT Innovation Centre
University of Southampton
Southampton, United Kingdom
{bmn, ve, sgc, pww}@it-innovation.soton.ac.uk

***Abstract*—Online communities play a pivotal role in innovation, marketing, corporate expertise management, product support and advertising. Communities in the order of millions of users are becoming the norm. However, this proliferation of demand is not met with intelligent, scalable, easy to use community management approaches. Current methods are based on basic statistical tools that aggregate data for the community owner/moderator to interpret and take appropriate actions. The data reflects only the current state of the community, which does not constitute an effective warning system of future events. Moreover, the community health becomes highly dependent on the owner's skill, interpretation, intimate knowledge of the community and its evolution path. This paper presents a proactive, extensible, risk-based management framework supporting advanced analytical services for managing online communities. The solution allows community owners to focus on the community objectives and proactively manage favourable/unfavourable events at the user and community level.**

***Index Terms*—Risk management; online communities; risks and opportunities; modelling; simulation; prediction.**

I. INTRODUCTION

Online communities generate major economic value and form pivotal parts of corporate expertise management, facilitating knowledge dissemination and communication as well as boosting performance and innovation [1], [2]. Research by McKinsey [3] shows that companies see a number of benefits in using collaborative technologies not only within their organisational boundaries but also for the purpose of reaching out to their customers, partners and suppliers. These advantages include faster access to knowledge and experts; increased customer and employee satisfaction; and a reduction of communication time and travel costs.

These findings are confirmed by further analysis by Deloitte and Frost & Sullivan [4], [5], [6], examining how social collaboration technologies impact business performance. Their research have shown that organisations who decide to deploy social networking tools within their organisational boundaries have much greater chance to improve their performance, attract customers, and establish profitable and long-term relations with them. Whilst there is a clear gain provided by such infrastructures, the management and preservation of their efficient operation is not trivial [7]. Communities can exceed millions of users and infrastructures must support hundreds of millions discussion threads that link together billions of posts. Current management solutions fail to meet current challenges of scale and growth, let alone the support for understanding and managing the business, social and economic objectives of the users and the host [1], [2], [8].

Current solutions usually consist of a dashboard for monitoring a set of Key Performance Indicators (KPIs, e.g., page views, number of posts, average time for responding/closing users' queries) judged relevant for the users' and the community's quality of service and experience. These KPIs present the state of the community at a certain point in time, current or past, but offer little to support what managers really need to effectively manage risks in the community; an insight onto the future state of the community. Community managers can be proactive if they have knowledge such as the likelihood that an expert's activity will drop within the next month and the impact on the community if this happens, the likelihood that they will miss their performance target of solving any query within 3 days, or the likelihood that negative sentiment will be developed on a certain topic. This information about the future state of the community is not provided by current management approaches. A successful community requires that the manager be empowered with new tools to break out of the current reactive framework by predicting the community future trends and the impact of any intervention across the community landscape as well as decide whether this is in line with the set of community objectives.

ROBUST [9] is an EC FP7-ICT project targeting an integrated, coherent view of the community dynamics in relation to the community objectives, offering a consistent, proactive approach to the management of the community. ROBUST is addressing the next generation of community management where the focus is not limited to managing failures but to manage risks and opportunities [7], [10], [11]. This paper presents the risk management work done in the ROBUST project, focusing on the risk management process applied to online communities and the architectural design of the proposed framework developed.

In Section II, we review the existing approaches for community management and their limitations. Section III present the risk management context from risk management standards perspective. Section IV details the risk specification work being done in the ROBUST project in the context of online communities. In Section V we present the risk management framework architecture. The conclusions and future work are

 ISBN: 978-1-61208-280-6

addressed in Section VI.

II. Managing Online Communities

Whilst the overall trend in adopting social technologies is positive, recent stories of BASF [12] and Alcatel-Lucent [13], who successfully deployed social networking technologies within their organisational boundaries, confirm that the benefits of these tools come at a cost. First, transforming the company into a social, collaborative network of employees, customers and partners is a disruptive step that breaks down traditional, hierarchical business models upon which a company relied (possibly from its inception). Second, the key to the successful transformation is not the technology itself (e.g., deploying wikis, social portals, messaging tools) but the understanding how this technology can be applied by people to achieve company-level goals (such as fast access to knowledge and experts, collaborative problem solving and cut in communication cost). This, in turn, not only requires clear understanding of the goals that such social systems need to fulfil, but more importantly, involves the continuous monitoring and steering of such systems to make sure they continue to satisfy company-level objectives.

This goes beyond the capabilities of information system management solutions [7] because of two main reasons: networked communities do not simply consist of machines and software, but are also communities of people and human behaviours; additionally, social software creates networked social environments in which individuals self-select whether and how they participate. The challenge in this ecosystem is to understand and manage the bidirectional relations governing how interactions between community members influence the overall community dynamics and impact its health and performance.

Given that online communities (and social media in general) is a relatively new phenomenon that emerged on a large scale a couple of years ago, we are just beginning to understand that the maintenance of such complex information ecosystems requires active management effort realised by highly skilled experts that take on the role of community managers. Unfortunately, for many first-time social media 'integration' efforts the official community management staff is often non-existent or represented by a few non-dedicated volunteers [14]. Tom Humbarger (Marketing and Social Media Strategy manager at AppleOne) is one of the few people who has done an analysis of community activity with and without a community manager. His observations [15] show that in the absence of community management, the activity (whilst not immediately terminal) slows significantly in a fairly short amount of time. The industry's reaction focused on the human aspect by allocating more community staff. For example, in the case of BASF, the active management has led to the involvement of a number of full-time community managers and part time involvement of staff (for governance) and solution stakeholders/owners. In addition, BASF utilises the power of numerous advocates who volunteer time to spread awareness and best practices across the company, as well as users who build and help facilitate communities of practice.

As social media adoption continues and thousands of online communities emerge, the role of community management matures and is perceived not only as a risk mitigator but also as a way to ensure that participation takes place, ROI is measured and business goals are being met [7], [10], [11]. Sophisticated data mining tools for sentiment and topic analysis are becoming essential in the monitoring platforms in order to provide more insights about the community [16], [17], [18]. However, these tools did not change the management methodology that is centered around monitoring the current state of the community, for instance, by looking at the daily growth rate of new members, calculating churn rates, identifying topics and sentiments, etc. Whilst this approach provides valuable information about the community's current or past state, it offers little support to the community manager in their endeavour to analyse the monitoring information and infer conclusions about the community's future. For example, it is hard for the community manager to predict the likelihood of a user developing negative sentiments or quantify the impact on the community caused by loss of key community members or a decrease in their activity. As a consequence, the correct choice of management actions is a mere trial and error process.

To address these limitations, this paper presents a novel risk-based approach for online community management in which decisions of community manager are supported by a suite of automated tools that constitute a community management platform. The goal of such a platform is not only to assist in the performance of tedious and time consuming routine tasks, such as monitoring of the current community state but, more importantly, to support pro-active community management involving the detection of risks and opportunities that may happen within the community in the near future. The key concepts behind the proposed solution along with its design architecture are presented in the remainder of this paper.

III. Risk Management

Many risk management methodologies can be found in the literature ranging from the generic [19] to the domain-specific [20] . The aim of this section is to shed light on the definitions of risk and risk management viewed by some of the highly acclaimed standards rather than offer an exhaustive list of them.

Management of Risk (M_o_R) [21] is a methodology published by the OGC (Office of Government Commerce). It defines risk as *"an uncertain event or set of events that, should it occur, will have an effect on the achievement of objectives. A risk is measured by a combination of the probability of a perceived threat or opportunity occurring and the magnitude of its impact on objectives"*. M_o_R considers threats to be the events with negative impact whereas opportunities are interpreted to have a positive impact.

The Risk Management standard [22] adopted by FERMA defines risk as *"the combination of the probability of an event and its consequences"*. It also notes the potential for events

and consequences that constitute opportunities for benefit (upside) or threats to success (downside).

ISO 31000 "Risk Management Principles and Guidelines" [19] defines risk as the *"effect of uncertainty on objectives ... An effect is a deviation from the expected — positive or negative"*. This definition agrees with the above descriptions on the uncertainty element of risk and its positive/negative relation to the objectives. However this controversial definition defines the risk as the impact (effect) of the uncertain event. ISO 31000 still refers to the traditional view of the risk being *"characterized by reference to potential events and consequences, or a combination of these"*. This ISO standard also indicates that *"risk is often expressed in terms of a combination of the consequences of an event and the associated likelihood of occurrence"*.

In ROBUST we specify a 'risk' in terms of an uncertain event (or set of events), which, if it occurs, affects the objectives of the community owner negatively; we reserve the 'opportunity' term for events that affect the objectives positively. It is possible for an event to affect multiple objectives negatively and positively at the same time. The boundary between risk and opportunity can get blurry, however, we argue that the community owner's decision on how to deal with the event ultimately classifies it as a risk or opportunity. This can be based on an implicit or explicit hierarchy of prioritised objectives.

The above standards have similar definitions of risk management being mainly the set of activities systematically applied to direct and control an organisation with regard to risk (identification, assessment and treatment). Risk management process, according to ISO 31000, consists of phases establishing the context, risk assessment (identification, analysis, and evaluation), risk treatment, communication and consultation, and monitoring and review - see Fig. 1.

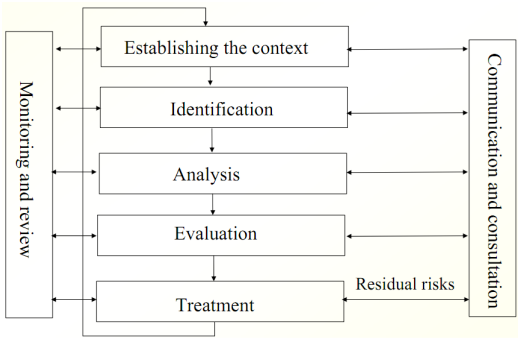


Figure 1. ISO 31000 risk management process

In the first phase, the context is specified, including the objectives and scope of the analysed system (online community in this case). Risk identification deals with the identification and specification of the risks and their attributes (e.g., events as well as their causes and potential consequences). Risk analysis involves developing a detailed understanding of the risk and determining the likelihood and consequences (i.e., level of risk) . A risk evaluation process classifies the risks according to acceptable risk criteria in order to make decisions about which risks need treatment and the priority for treatment implementation. The treatment options could be to avoid the activity that gives rise to the risk, modifying the risk likelihood or consequence (enforcing countermeasures), retaining/accepting the risk or sharing it with another party. ISO 31000 does not distinguish treatment actions for a risk from those of an opportunity. However, M_o_R offers more specific options shown in Table I, below:

Table I
RISK AND OPPORTUNITY RESPONSES

Risk response	Opportunity response
Avoid	Exploit (ensure that event occurs and impact realised)
Reduce (likelihood or impact) Fallback (reduce impact) Transfer to 3rd party (e.g., insurance)	Enhance (likelihood or impact)
Share	
Accept	Reject (no action)

The ROBUST framework supports an organisation's risk and opportunity management activities by providing tools that assist in risk identification, expression, analysis, evaluation and decision making in the treatment process. The tools do not mandate the adoption of any specific risk management process. The details of the framework are presented in the following sections focusing on how ROBUST addresses each of these phases.

IV. ONLINE COMMUNITY RISK SPECIFICATION

A. Risk Components

As per the definitions in Section III, risks and opportunities are defined within the context of the organisation and its objectives. These objectives may be strategic, tactical or operational. The scope can be the whole organisation, a department or a sub-system such as the online community. The online community can have its own objectives (e.g., knowledge transfer between users, improved quality of experience, increase the number of community members) or inherit the organisations objectives (e.g., reduce operational costs).

The objectives can then be used in order to identify the uncertain events that may affect them positively or negatively. The objectives can be viewed as the scope for the process of identification of risks and opportunities. An event is the (one or more) occurrences or non-occurrences of a change of a particular set of circumstances. An event can lead to a range of (certain or uncertain) impacts on objectives. The impact level or severity may not be static and may vary according to different factors (time, life-cycle phase, community features, etc.). An impact scale is usually produced in order to quantify the risk impact (e.g., high, medium, low). The scale should be specified in terms of impact criteria in line with the

 ISBN: 978-1-61208-280-6

objectives (e.g., in terms of activity level drop, content quality deterioration or financial loss).

B. Event Specification

In ROBUST we classify risk/opportunity events based on their source being either internal or external. The external events are those that originate from external actors whom are not part of the community. Examples of external events include the introduction of new legislation affecting the online community or the launch of a competitor's community leading to the churn of users.

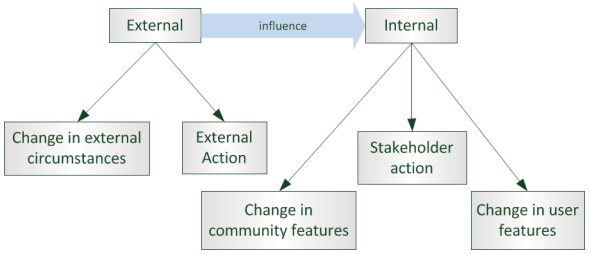


Figure 2. Event categories

The internal event category comprises those events that originate from within the community, such as a modification of its structure or a change of an individual user's attributes (see Fig. 2). The events can be deliberate, accidental or the result of the normal evolution of the community. The internal events can be decomposed into three categories:

1) Community features: This includes any changes related to the community attributes like content, structure, users, performance, etc (e.g., drop in community members or new joiners)
2) User features: This includes any changes related to the user attributes including role, connections, position (e.g., change of role, network centrality or activity level)
3) Stakeholder(s) actions: this includes deliberate or accidental actions performed by the community managers or users (e.g., delete or block user, user flaming others)

Note that the three events categories are inter-dependent since actions by stakeholders influence community features as well as user features. The same applies to the external events that can influence the internal changes. Moreover, within the same category e.g., community features category, changes in the number of community users may influence the community activity level. In ROBUST, we focus on the internal events mined from community data logs that contains users' activities time series. Given the event definitions we have established, events can then be formulated as: Variable(s) attaining a certain value (i.e., threshold if numeric) or Variable(s) changing from one value to another (i.e., states).

An example of the former is 'the drop in community members exceeding 50%', and for the latter, 'a user X changing their role in the community from *contributor* to *lurker*'. Although these are examples of risks formed by a single event, there can also be compositions of multiple variables to produce more complex events. For instance the community owner may be interested in the joint event of an "expert user change of role from contributor to lurker" and "increase in community new joiners" in order to anticipate the load.

With risks/opportunities specified in such a measurable manner, it is possible to automatically perform monitoring and assessment in order to compute the likelihood of the events, analyse them and later decide on a treatment strategy. In the following Section, we introduce the ROBUST Risk Management Framework and how it supports the risk management cycle.

V. The ROBUST Risk Management Framework Implementation

The risk management framework that is presented here takes a central role in a platform that is produced in the ROBUST project. It is designed to be flexible, extensible and to allow the integration of new modules to address a multitude of risks and opportunities in different communities. An overview of the framework is given in Section V-A, followed by sections describing how the framework can be used in the risk management process discussed in Section III for managing online communities. For this, a examples based on the SAP Community Network will be given [23].

A. Overview

An overview of the main components in the framework is given below in Fig. 3, comprising three layers: *presentation*, *core* and *support* layers.

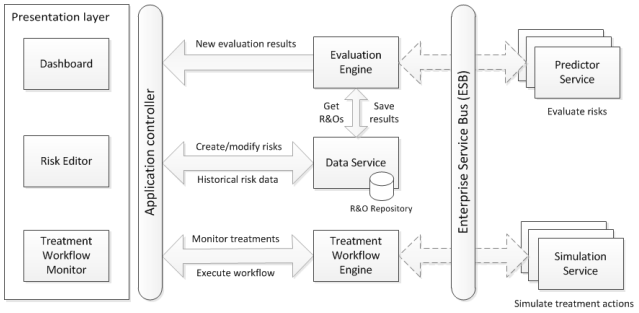


Figure 3. Risk management framework components

The *presentation* layer mainly includes the interfaces available to the community manager to interact with the system capabilities. It consists of a Risk Editor, a Dashboard that incorporates the results of the risk analysis and visualisation tools, and treatment workflow monitor to manage the management response to risks and opportunities.

The *core* layer consists of a Risk Registry, a Workflow Engine that enacts the treatment workflows and an Evaluation Engine that orchestrates support layer services in order to evaluate the risk and opportunity events.

In the *support* layer, different services may be made available that can perform analytical tasks. We distinguish between two types of services:

 ISBN: 978-1-61208-280-6

- Predictor Service: provides event prediction capabilities based on the community data (could be done via batch processing of historical data or real-time stream processing).
- Simulation Service: allows simulating the community evolution under a variety of "what if" scenarios and, thus, enabling the assessment of the impact of the different events and actions.

As illustrated in Fig. 3, above, the support services are connected to the core system via an Enterprise Service Bus (ESB) to improve the robustness and extensibility of the framework. This means any number of services can be made available for bespoke communities, exposing different functionalities to address any risks or opportunities identified for the respective community. To validate the ideas described in this paper, we developed a prototype of the proposed risk management framework based on Java technologies using Apache ServiceMix ESB as a backbone. The prototype was applied on one of the ROBUST use cases: the SAP Community Network (SCN) [23]. This is an online platform of multiple communities for customers seeking support about SAP products. Community members (whether SAP employees or not) are encouraged to contribute to the community by being awarded points for answering other users' questions. The following sections describe the implemented functionality and interactions supporting the main risk management phases discussed in Section III taking as example the risk of missing a performance target: solving any customer query within 3 days.

B. Establishing the Context

The framework provides plugins to visualise and analyse online communities as part of establishing the context (roles, topics, network). The manager provides the system with details of the sources of community data, as data batches or real-time streams, and specifies the community objectives. In our case, this included, *inter alia*, quality of experience and performance, fostering sharing and healthy community growth [24].

C. Identification and Specification

The identification phase is based on the community objectives and manager's knowledge of their own community events and impacts. ROBUST tries to minimise the latter dependency via Simulation Services that allow exploring the future state of the community in various "what if" scenarios. The manager can then identify the relevant events and can proceed with the specification using the Risk Editor.

The specification includes the risk (opportunity) title, owner (who is responsible for this risk management), scope (one user, group of users or community) as well as the risk event(s) and impacts (on the objectives specified above). In our scenario, the event "not providing customer support within a reasonable time" was identified as a risk. This is expressed in a measurable form of a ROBUST risk as "thread response time exceeds a threshold t".

The information about supported events originates from the Predictor Services, each of which supports one or more event prediction capabilities. We have developed a Predictor Service that can estimate the likelihood of the above event using agent-based modelling (details are out of scope in this paper). Advanced tools developed in the ROBUST project are being integrated to address events based on topics, sentiment and quality of the content [25], as well as the roles and behaviour of users [26].

D. Assessment

This phase entails design-time as well as run-time activities. In the design-time, the community manager needs to develop the understanding of the event by quantifying any variables like the above threshold, t. This can be guided by the manager's own observations/knowledge, existing online communities' best practices or inherited from the organisation's policy as it was in our example (3 days). Moreover, the event impact on the objectives is also defined in this phase by indicating the relative potential impact (positive or negative) the event may cause, should it occur. In our example, we can estimate a negative impact on all objectives specified above, particularly on high performance and quality of experience objectives [24].

The above specification is stored in the Risk Repository and used at run-time by the Evaluation Engine (EE). The EE orchestrates the associated Predictor Service to evaluate the event likelihood within a future time window, such as the next week or month. The future time window predictions can be made for depends on the capabilities of the respective Predictor Service. Prediction capabilities of the system can always be extended by implementing a Predictor Service API, which includes a generic way of describing the supported events and any required configuration parameters.

The results of the Predictor Services are returned asynchronously to the EE, which stores the results in the Risk Repository and notifies the Risk Dashboard to give real-time updates to the risk manager. An example of a visualisation provided in the Dashboard is a risk/opportunity matrix, which shows the likelihood of the risk on the vertical axis and its impact (positive or negative) on the horizontal axis. The manager can then evaluate the risk and opportunity levels and proceed with the treatment phase if required.

E. Treatment

The Risk Editor allows the manager to assign a treatment workflow as a response (mitigate, fallback or exploit) for the risk or opportunity during design-time and enacting it during run-time. Treatments are workflow based plans that address one or more risks (or opportunities). Each treatment plan includes a linear or parallel series of actions that can (some could be optional) be carried out by the risk/opportunity owner. Whilst some of these actions may require direct and manual intervention in the community, such as blocking a malicious user, others may dictate community exploration using other ROBUST tools, for instance, using a Simulation Service to explore and evaluate the potential outcomes of planned

 ISBN: 978-1-61208-280-6

actions. The Treatment Workflow Monitor implemented in our framework allows storing, retrieving and enacting BPMN 2.0 based workflows to support semi-automated responses based on the open source Activiti API [27]. In our example a simple treatment workflow can be based on "allocating additional internal resources, or attracting external resources by providing double points for answering questions during the next period". The impact of these actions can be evaluated, before actual enactment, using our "what-if" Simulation Service by adding more agents or changing the agents' behaviour.

VI. Conclusions and Further Work

This paper has reviewed the current practices of community management and showed the need for a new generation of tools focusing on proactive risk management. We have presented a framework that goes beyond analysing the current state of the community; focusing on predicting whether risks or opportunities are likely to occur in the future and what their impacts may be on the community.

ROBUST supports online community risk and opportunity identification, specification, assessment, monitoring, visualisation and treatment. To our knowledge, such a risk-based approach has not been directly exploited for online community management. In our initial evaluation with SAP and IBM community owners, it was pointed out that risk management expertise can be an obstacle for the approach acceptance. However, the promising capabilities provided by this approach justifies the minimal effort of acquiring the expertise. Though scalability of the framework was addressed by design (modular, loosely coupled services), the accuracy of predictions depend on the predictor services of which we developed Gibbs Sampler, Compartment Model, and Agent-Based Simulation Service [9]. The loosely coupled design guarantees that other statistical approaches can be integrated if needed.

The prototype framework currently implemented does not model the interdependencies between risks. This is not a trivial problem that we are intending to tackle in our future work.

Portability of the system is crucial and requires supporting different online community data sources and models. In ROBUST, a common community data model based on SIOC [28] is being developed for this purpose. We are in the process of updating our prototype to support the new community data model thus allowing any community to use the framework by simply mapping their data schema to SIOC.

Acknowledgements

This work has been carried out in ROBUST, an EC supported 7th Framework Programme ICT project (FP7-257859).

References

[1] A. Mocan, F. Brauer, and W. Barczynski, "D8.1: Provisioning and Preparation of the SAP Community Network Data," EC FP7-ICT ROBUST Project, 2011, http://www.robust-project.eu/results/provisioning-and-preparation-of-the-sap-community-network-data/view [retrieved: April 2013].

[2] I. Ronen, S. Ur, and I. Guy, "D7.1: IBM Employee Network Data and Requirements," EC FP7-ICT ROBUST Project, 2011, http://www.robust-project.eu/results/ibm-employee-network-data-and-requirements/view [retrieved: April 2013].

[3] McKinsey, "The rise of the networked enterprise: Web 2.0 finds its payday," http://www.mckinseyquarterly.com/The_rise_of_the_networked_enterprise_Web_20_finds_its_payday_2716, [retrieved: April 2013].

[4] Frost & Sullivan, "Meetings Around the World: The Impact of Collaboration on Business Performance," http://www.jmorganmarketing.com/wp-content/uploads/2010/03/impactcollab.pdf, [retrieved: April 2013].

[5] ——, "Meetings Around the World II: The Impact of Collaboration on Business Performance," http://www.jmorganmarketing.com/wp-content/uploads/2010/03/impactcollab.pdf, [retrieved: April 2013].

[6] Deloitte, "Social software for business performance." http://www.deloitte.com/assets/Dcom-UnitedStates/Local%20Assets/Documents/TMT_us_tmt/us_tmt_%20Social%20Software%20for%20Business_031011.pdf, [retrieved: April 2013].

[7] R. Happe, "Announcing the 2012 State of Community Management Report," http://www.slideshare.net/rhappe/2012-state-of-community-management-12162160, [retrieved: April 2013].

[8] M. Boniface, J. Pickering, E. Meyer, C. Cobo, A. Oostveen, B. Stiller, and M. Waldburger. (2011) D3.1 First Report on Social Future Internet Coordination Activities. EC FP7-ICT SESERV Project. [Online]. Available: http://www.scribd.com/doc/68338983/D3-1-v1-5

[9] EC FP7-ICT ROBUST Project, http://www.robust-project.eu/, [retrieved: April 2013].

[10] R. Happe, "The 2010 State of Community Management Report. Best Practices from Community Practitioners," http://community-roundtable.com/socm-2010/, [retrieved: April 2013].

[11] ——, "Announcing the 2011 State of Community Management Report," http://community-roundtable.com/socm-2011/, [retrieved: April 2013].

[12] D. Hinchcliffe, "Enterprise 2.0 success: BASF," http://www.zdnet.com/blog/hinchcliffe/enterprise-20-success-basf/1939?tag=search-results-rivers;item19, [retrieved: April 2013].

[13] ——, "Enterprise 2.0 Success: Alcatel-Lucent," http://www.zdnet.com/blog/hinchcliffe/enterprise-20-success-alcatel-lucent/1917?tag=content;siu-container, [retrieved: April 2013].

[14] ——, "Community management: The essential capability of successful Enterprise 2.0 efforts," http://www.zdnet.com/blog/hinchcliffe/community-management-the-essential-capability-of-successful-enterprise-20-efforts/913, [retrieved: April 2013].

[15] T. Humbarger, "The Importance of Active Community Management Proved With Real Data," http://tomhumbarger.wordpress.com/2009/01/13/the-importance-of-active-community-management-proved-with-real-data, [retrieved: April 2013].

[16] Simply Measured, http://simplymeasured.com/, [retrieved: April 2013].

[17] Sysomos, http://www.sysomos.com/, [retrieved: April 2013].

[18] Facebook Insights, http://www.sysomos.com/, [retrieved: April 2013].

[19] ISO/IEC, *31000:2009 Risk management - Principles and guidelines*, ISO Std., 2009.

[20] ——, *ISO/IEC 27005:2011 Information technology - Security techniques - Information security risk management*, ISO Std., 2011.

[21] Office of Government Commerce (OGC), *Management of Risk: Guidance for Practitioners*, Office of Government Commerce (OGC) Std., 2010.

[22] IRM, *A Risk Management Standard*, IRM Std., 2002.

[23] SAP, "SAP Community Network," http://scn.sap.com/, [retrieved: April 2013].

[24] A. Mocan, F. Brauer, and B. Massarczyk, "D8.2: Business Analysis of the SAP Ecosystem," EC FP7-ICT ROBUST Project, 2011, http://www.robust-project.eu/results/business-analysis-of-the-sap-ecosystem/view [retrieved: April 2013].

[25] C. Lin, Y. He, R. Everson, and S. Rueger, "Weakly-supervised joint sentiment-topic detection from text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 1134–1145, 2012.

[26] S. Angeletou, M. Rowe, and H. Alani, "Modelling and Analysis of User Behaviour in Online Communities," in *Proc. of the International Semantic Web Conference*, 2011, pp. 35–50.

[27] Activiti Business Process Management platform, http://www.activiti.org/, [retrieved: April 2013].

[28] SIOC, "Semantically-Interlinked Online Communities," http://sioc-project.org/, [retrieved: April 2013].

 ISBN: 978-1-61208-280-6

Enforcing Data Availability in Structured *Peer-to-Peer* Storage Systems *With Zero* Replica Migration

Mesaac Makpangou
REGAL Team
INRIA/LIP6 (UPMC)
Paris, France
Email: mesaac.makpangou@lip6.fr

***Abstract*—This paper presents a structured peer-to-peer storage substrate that exploits notifications issued by the underlying network maintenance layer to enforce data availability, while avoiding both application-level replica tracking and unneeded replica migrations. This system enforces a multiple keys replication approach. Each peer advertises its stored contents to a set of watchers picked from the set of peers within this peer's neighborhood. When a peer departs from the overlay network, its watchers initiate repairs of losses due to this departure. Thanks to the location of watchers within each watched peer's neighborhood, on peer departure, one can reduce the overall loss repair delay, and hence the probability of loosing for ever a stored content. The analytical evaluation shows that the proposed replica maintenance substrate generates far less overhead than a leaf set based replica maintenance system. Furthermore, on node arrivals, the overhead incurred by this proposal does not depend on the size of stored contents. This makes this substrate an interesting building block for peer-to-peer storage systems destined to store large-size objects.**

***Index Terms*—Peer-to-peer storage system, replica maintenance, flexible replication, distributed algorithms.**

I. Introduction

One challenge for a structured peer-to-peer storage system is to efficiently enforce stored contents availability in the face of node churn. One well known technique to address this issue is data replication. Number of existing structured storage systems enforce the leaf set (or successor) based replication approach. Each data item is replicated at its replica set (i.e., the k closest nodes to its root node, where k is the replication degree enforced by the system). In practice, upon each node departure, a replica maintenance procedure is run to create replicas that are lost due to this departure. Also when a node joins the overlay, it cooperates with its peers to determine replicas to migrate at the new coming node. While this solution to maintain data availability is simple, the overhead due to replica migrations increases with the number and the total size of stored contents [1]. This could jeopardize the overall performance of the system, especially if we consider the replication of large-size objects.

To enforce replica availability while avoiding unneeded replica migrations, one approach is to use multiple publication keys and to let the storage system computes the suitable number of storage keys, then uses them to place object replicas within the storage overlay network. This approach requires a means to efficiently track the availability of individual replicas. One solution is to check periodically the availability of each replica and to recreate a replica only when a replica is really lost. Such an active tracking of replicas has two drawbacks. Firstly, the loss repair delay is likely too high: to limit the tracking cost (i.e, number of messages and consumed bandwidth), one tends to enforce a large probe period; unfortunately, the larger the probe period, the higher the replica loss repair time and consequently the higher the risk to lose forever a stored data [2], [3]. Secondly, the active tracking doesn't scale: as the number of replicas augments, the tracking cost augments too; at some point, there will be not enough resource left for useful work.

This paper presents a Peer-to-peer Watching System (*Pws*), a replica maintenance substrate for structured storage utilities. *Pws* relies on a distributed hash table (DHT) abstraction; existing implementation of this abstraction as proposed in [4]–[7]. *Pws* enforces the multiple publication key replication approach. It associates with each storing node a *watch set*, a subset of nodes located within this node's neighborhood. Each storing node advertises to its watchers (i.e., members of its watch set) each object (i.e., replica or manager) that it stores, together with the identifier of the group of managers capable to handle this object's loss. Whenever a watcher detects that the node that it is watching departed (voluntarily or involuntarily) from the storage overlay network, it notifies manager groups of objects stored at that node to repair their losses. *Pws* is layered on top of an underlying network overlay that provides the common application programming interface specified in [8]. In particular, *Pws* assumes that the underlying network management system notifies each node whenever an event (i.e., node departure or arrival) that impacts this node's neighborhood occurs. *Pws* exploits these already existing notifications to track replica availability and to ensure that each node's watch set members are within this node neighborhood. Thanks to both the location of watchers within each node's neighborhood and notifications of changes in the neighborhood by the underlying network management layer, *Pws* detects

 ISBN: 978-1-61208-280-6

node departure (and hence stored contents loss) with (almost) none delay. This contributes to reduce the replica repair time which is a key metric that impacts the data availability in peer-to-peer storage systems [2].

The contribution of this paper is twofold. Firstly, a replica maintenance substrate that permits to build structured peer-to-peer storage systems that enforce data availability without incurring the cost of unneeded creations of replicas and of a periodic replica tracking mechanism, while enabling the detection of replica loss with almost no delay. Secondly, a flexible replication model that separates replication concerns: number of replicas, replica loss repair strategy, replica placement, replica watching. Applications control the replication scheme and the replica repair strategy, while the system is responsible of locating and watching alive replicas. This model permits providers to adapt the replica maintenance strategy, and hence the quality of service offered to end-users, on a per object basis. A detailed description of the main algorithms, together with the evaluation of the system overhead are provided.

The rest of the document is organized as follows. Section II presents related work, while Section III gives an overview of the *Pws* system. Then Section IV details the main algorithms and protocols of *Pws*. Section V evaluates *Pws* cost and compares it to a basic leaf set-based replica maintenance. Finally, Section VI draws some concluding remarks.

II. Related Work

Several neighborhood-based replication algorithms have been proposed to ensure data availability in structured peer-to-peer storage systems. For these systems, each data is stored at its root node and at a subset of neighbors. For instance, in PAST [9], a large-scale persistent storage utility using Pastry [6], the replicas of a file are stored in the k nodes that are numerically closest to the file identifier, where k is the replication degree to enforce [10]. One main advantage of the neighborhood-based replication is its capacity to efficiently tolerate node failure. When a node fails, a node in the neighborhood of the failed node is automatically promoted the responsible of data that were stored at that node and will be the target of lookup requests. While this solution to maintain data availability is simple and transparent to applications, the overhead incurred to create new replicas on replica set changes is unacceptable for storage systems that experience high churn. RelaxDHT [11] proposes to relax the replica placement constraint such as to avoid creating replicas when this is not mandatory to preserve data availability. For each stored object, replicas can be anywhere in the leaf-set (i.e, not necessary at the k closest nodes to the root node); however, the root node maintains meta-data describing its replica set. The root node periodically, sends messages to the replica set peers so that they keep storing their replicas. While this relaxation constitutes a real improvement, it still incurs unneeded replica creations when a member of a replica-set is put out of the root node's neighborhood, following node joins. *Pws* extends this relaxation and enforces a complete separation of a number of concerns that are often tightly coupled: replica placement, replica location, and replica maintenance. Thanks to this separation, *Pws* proposes a scalable way to track replicas of each object scattered over the network, together with a flexible replication management that permits to adapt the replication strategy (number of replicas, placements) according to the quality of service required by the object provider.

A. Ghdsi et al. [12] propose a replication scheme called symmetric replication. Each identifier is associated with k equivalent identifiers, where k is the suitable replication degree. That is, if N is the set of identifiers, N is partitioned into N/k classes. Each object is replicated at the nodes whose identifiers are equivalent to the root node's identifier. To preserve this invariant, when a new node joins the system, it cooperates with members of its class to obtain the list of objects to replicate. To access an object, one addresses the lookup request to any peer that stores the object replica, that is any node that belongs to the same class as the object root node.

Total Recall [2] is a peer-to-peer storage system that implements data availability as a first class storage property. It provides a means to characterize host availability. Total Recall proposes an architecture that permits to adapt the redundancy mechanism in function of the host availability characteristics. In practice, the system continuously monitors its constituent hosts, then derives the average host availability. Given a host availability level and a specified availability target, Total Recall provides insights on how to determine the redundancy degree that ensures with high probability that the data remains available. Total Recall permits to adapt the redundancy mechanism to the specifics of data and/or of the hosting infrastructure. Unlike Recall, *Pws* focuses only on ensuring replica availability.

III. System Overview

From the user point of view, *Pws* is a DHT-based storage system. Each replicated object is associated with a key that designates the group of replicas of that object. From the system point of view, *Pws* implements a flexible replicated object model, a multiple publication key replication approach and a neighborhood-based tracking of replica availability. It can serve as a replica maintenance substrate for peer-to-peer storage utilities destined to serve large size objects. An effort to develop such a system is presented in [13].

A. Flexible Replicated Object Model

Each object is associated with a replication contract that controls mainly the number of replicas to create and the replica loss repair strategy. A replication contract is enforced thanks to the cooperation of a group of managers. There is one manager group per replicated object.

Overall, within *Pws*, a replicated object is represented by two sets: R, the set of object replicas; and M, the set of replication managers that cooperate to enforce the replication contract associated with this object. Both object replicas and their associated replication managers are stored within the underlying storage overlay network.

 ISBN: 978-1-61208-280-6

The group of managers controlling a replicated object, as a whole, has the responsibility to maintain the suitable number of object replicas and of replication managers. These numbers may vary from one replicated object to another. It is worth noting that, for a given replicated object, its managers repair both object replica and manager loss.

B. Replica Placement

To replicate an object, *Pws* proceeds in three steps. Firstly, it determines the number of replicas to create and the number of managers that will be in charge of managing this replicated object, then attaches one distinguished local identifier to each replica and to each manager. Note that these decisions are guided by the replication contract associated with the object provider. A basic replication contract simply indicates the minimum number of replicas and of managers that the system has to maintain alive.

Secondly, *Pws* computes a distinguished key for each manager, then uses the computed key to place the corresponding manager. Upon the completion of each manager's installation, *Pws* notifies its identifier and its current location to the group of managers controlling the same replicated object.

Thirdly, once all managers of an object are placed, *Pws* places the object replicas. Again, for each replica, *Pws* computes a distinguished key, then uses it to place this replica. Once a replica is stored, *Pws* notifies its current location to managers controlling this replicated object.

Overall, upon the completion of the replication procedure, the suitable number of replicas and of managers are stored within the storage overlay network. Each manager of the new replicated object is aware of the current locations of all object replicas and of all managers.

Pws enforces final replica placements. That is, once a replica or a manger is placed at one node at the replication time, it will remain stored there as long as that node remains alive, regardless arrivals or departures of other storing nodes.

C. Neighborhood-based Peer's Availability Monitoring

To maintain the availability of a replicated object, *Pws* has to ensure that at anytime at least one replication manager and one replica remain available. *Pws* relies on watchers within each storing node's neighborhood to track the availability of both replicas and mangers stored at each storing node.

In practice, *Pws* associates with each storing node a set of nodes located within this node's neighborhood, called its *watch set*. *Pws* guarantees that, at each time, members of a node's watch set are all within this node neighborhood. This guarantees that when a node fails, its watchers are promptly informed by the network maintenance layer.

Each storing node advertises its contents to its watchers. An replica advertisement carries in particular the replica identifier together with the manager group identifier of the concerned replicated object. Thanks to these advertisements, a watcher can determine the contents lost when it is informed that a node that it monitors departed from the overlay network and can subsequently notify each loss to managers in charge of repairing this loss.

Overall, the neighborhood-based peer's availability monitoring permits to detect replica loss and to request its repair with no delay. This contributes to reduce the replica repair time which is a key metric that impacts data availability in a peer-to-peer storage system [2].

IV. Peer Watching System

The primary role of the peer watching system (Pws) is to maintain replica availability in presence of node churn. The key idea is to avoid replica migrations while enabling to detect replicas' loss with no delay and without paying for active replica tracking. For that, each storing node is associated with a set of nodes located within its neighborhood, called its *watch set*. Whenever a storing node leaves (voluntarily or involuntarily) the overlay network, its watchers repair its loss by notifying its departure to managers that control its stored contents.

A. Watch Set Initialization

A *Pws* node creates and initializes its watch set whenever a replica is first stored at this node since the last re-start. This is done in three steps.

Firstly, at the storing node, *Pws* requests the list of the node's replica set, thanks to an invocation to the primitive *replicaSet()* provided by the underlying KBR layer [8]. Once the replica set is retrieved, *Pws* selects at most $maxW$ elements to act as watchers of this storing node, where $maxW$ is a system configuration parameter, then sends a watch request message to each selected neighbor (see Function `createWatchSet()` of Figure 1). This message carries the identifier of the node to watch (i.e., the requester), the current value of the requester's `localTime`, and the set of selected watchers.

Secondly, at each selected watcher's node, upon the reception of a request to watch, *Pws* checks whether the local node has been watching this requester in the past (see Function `onReceiveRequestToWatch()` of Figure 1). If the receiver holds none information, it creates a new watched node descriptor, passing to the constructor the information contained in the watch request message. If however the receiver node does have a descriptor representing the requester within its `watchedNodes` structure, it updates that descriptor. Note that, this could happen if the receiver was a member of the requester's watch set, then at some point was pushed out the requester's neighborhood. Upon the completion of the treatment of the request to watch a server, the receiver acknowledged the highest timestamps received so far.

Finally, at the requester storing node, upon the reception of a response from each candidate watcher, *Pws* updates the structure that describes the corresponding watcher. In particular, for each watcher which is aware of replicas stored by this replica server, it registers the timestamps of the most recent received advertisement.

 ISBN: 978-1-61208-280-6

B. Watch Set Maintenance Protocol

The objective of the watch set maintenance protocol is to make sure that for each storing node, there always exists at least one watcher within this storing node's neighborhood. Though initially each watcher is peaked from the neighborhood of the storing node that it is watching, arrivals of new nodes can push it out this neighborhood. Also, watchers can leave (voluntarily or involuntarily) the overlay network.

Function `onJoinOrLeave()` of Figure 1 sketches the watch set maintenance procedure. The watch set maintenance procedure exploits update upcalls from the underlying overlay network maintenance facility. Precisely, when a change occurs in the local node neighborhood, an update upcall is issued to this node, passing it the identifier of the node that departed or joined [8]. The *Pws* provided `update()` then invokes the watch set maintenance procedure which behaves as follows. Firstly, *Pws* retrieves the list of neighbors of the local node then computes, w, the number of alive watchers that are still in the local node's neighborhood. If $w \leq wThreshold$, *Pws* selects $maxW - w$ neighbors that are not currently watching the local node and adds them to the local node's watch set. Secondly, *Pws* notifies the current watch set membership to watch set members.

C. Advertisement Protocol

The objective of this protocol is to make sure that, whenever a storing node leaves the storage overlay network, its watchers are aware of replicas stored there and of managers that can repair their losses. The problem arises from the fact that the contents of each storing node change over time: new replicas can be added and old replica can be deleted. The challenge is to conciliate conflicting requirements: maintaining the network bandwidth consumption as low as possible while ensuring that in case of the departure of a storing node, that node's watchers have an accurate list of replicas that are lost.

Periodically and after each reconfiguration of the watch set of a node, *Pws* requests the watching service to advertise the local node's contents to its watch set, thanks to the execution of Function `onRequestToAdvertise()` of Figure 2 that permits to advertise recently stored contents to members of the local node's watch set. An advertisement message contains mainly : the node identifier, the current local time, and the descriptors of replicas stored since the most recent advertisement that has been already acknowledged by all watchers. Each advertised replica descriptor comprises this replica local identifier, the group identifier of its managers, and its storage time.

Function `onAdvertisementDelivery()` of Figure 2 sketches actions performed by each watcher upon the reception of an advertisement. Firstly, it updates accordingly its view of the watched node. In particular, each watcher registers the descriptors of replicas that it was not yet aware of; it also updates the most recent timestamps received from the current sender. Then, upon the completion of its treatment, it acknowledges the reception and the treatment of the advertisement.

```
// Data structures maintained by each PWS node
thisNode; // Reference to the local node
watchSet; // Set of watchers of this node
wThreshold; // Minimum number of watchers per node
maxW; // Maximum number of watchers per node.
cts; // Current timestamps of the local storing node.
watchedNodes; // PWS nodes watched by this node

function createWatchSet()
Node[] replicaSet = thisNode.replicaSet();
if replicaSet.isEmpty() then
    return 0;
end if
for i = 1 to i = maxW do
    watcher = new Watcher(replicaSet[i]);
    watchSet.add(watcher);
end for
cts = initTimestamp();
notifyWatchers();
return watchSet.size();
end function

function notifyWatchers()
watchers = watchSet.getWatchers();
req = new RequestToWatch(thisNode.id, watchers , cts);
for all w ∈ watchers do
    route(w.getNodeId(), req, null);
end for
nbResponses = 0;
while nbResponses < watchers.size() do
    waitNextResponse2RequestToWatch();
    r = getNextResponse2RequestToWatch();
    watchSet.updateLastRcvTS(r.watcher, r.lastRcvTS);
    nbResponses++;
end while
end function

function onReceiveRequestToWatch(req)
n = req.requester;
if (watchedNodes == ∅)||(n ∉ watchedNodes) then
    watchedNode = new WatchedNodeDesc(n)
    watchedNodes.add(watchedNode);
end if
watchedNodes.updateInfo(n, req);
lastRcvTS = watchedNodes.getLastRcvTS(n);
r = new Response2RequestToWatch(thiNode.id, lastRcvTS);
send(n, r);
end function

function onNeighborJoinLeaveNotification()
Node[] neighbors = thisNode.neighborSet();
currentWatchers = watchSet.getWatchers();
for all w ∈ currentWatchers do
    if w ∉ neighbors then
        watchSet.remove(w);
        sz = watchSet.size();
    end if
end for
if (sz > wThreshold) then
    return
end if
replicaSet = thisNode.replicaSet();
while ((0 < i < maxW) && (sz < maxW)) do
    if (replicaSet[i] ∉ watchSet) then
        watchSet.add(replicaSet[i]);
        sz++;
    end if
    i++;
end while
notifyWatchers();
end function
```

Fig. 1. WatchSet membership maintenance

At the storing node side, upon the reception of a response from a watcher, *Pws* updates the corresponding state maintains within the `watchSet` data structure.

 ISBN: 978-1-61208-280-6

D. Replica Loss Detection and Notification

Given the set of watchers of a node, we define the primary watcher for this node to be the watcher with the smallest node identifier, bigger or equal to this node's identifier. When a node detects the departure of a storing node that it is watching, it first determines whether it is the primary or a secondary watcher of this storing node. As watch sets change over time and since changes are not notified atomically to watch sets' members, the watch set membership view can differ from one member to another. Hence, two distinct watchers can decide different primary watchers. This will result in more notifications than necessary.

If a watcher decides to be the primary watcher, it notifies the loss of each advertised replica to the advertised managers. Otherwise, it sends a request to participate to notifications to the watcher that it considers to be the primary. Such a request contains in particular the identifier of the node that has failed.

Upon the reception of a request to participate to notifications, the (presumably primary) watcher checks if it has already detected the notified departure and if it has decided to be a primary watcher of the failed node. If both conditions are not satisfied, it promotes itself a primary watcher for the departed node, then acts accordingly. Once notifications have been sent to suitable managers, it sends back a response to the requesting secondary watcher. This response contains the timestamps of the most recent replica advertised to it by the failed storing node.

Upon the reception of a response from a primary watcher, a secondary watcher checks it it were advertised replicas that its primary is missing. If any, it notifies their loss to their advertised managers. If however a secondary watcher receives none response to its request to participate to notifications, after some delay, it promotes itself a primary and acts accordingly.

Regardless which watcher (primary or secondary) does it, a replica loss notification contains the lost replica's identifier, its location, and its creation timestamps.

E. Replica Maintenance Protocols

Pws ensures data replica availability thanks to the cooperation between watchers in the neighborhood of storing nodes and managers of replicated objects, also replicated towards the overlay network. Watchers are informed with no delay of replica or manager losses, thanks to update upcalls issued by the overlay network maintenance layer. Once informed, watchers cooperate with one another to notify appropriate managers. Managers, in turn, guarantee that at anytime, at least one data replica and one manager remain available.

Pws balances the handling of replica losses among managers controlling each replicated data. In practice, the loss of a replica (or manager) is handled by the manager with the smallest local identifier among managers that remain alive, equal to or greater than the lost replica's local identifier. Each manager maintains the list of its peers that are still alive, and the list of outstanding repair requests.

Pws enforces two separate replica maintenance protocols: one for replication managers and one for object replicas.

```
// Additional structure maintained by each PWS node
storedReplicas; // Set of descriptors of replicas stored locally

function onRequestToAdvertise()
Timestamps lbackts = watchSet.computeLowerBoundAcknowledgedTS();
if lbackts < cts then
    adv = new Advertisement(thisNode.id, cts);
    for all r ∈ storedReplicas do
        if (r.storatgeTime > lbackts) then
            adv.addReplica(r.id, r.manager, r.storageTime);
        end if
    end for
    for all w ∈ watchSet do
        route(w.getNodeId(), adv, null);
    end for
end if
end function

function onAdvertisementDelivery(adv)
TimeStamp lastRcvTs = watchedNodes[adv.notifier].lastRcvTs;
if (adv.cts < lastRcvTs) then
    return lastRcvTs
end if
for all replica ∈ adv do
    if (replica.storageTime > lastRcvTs) then
        watchedNodes[adv.notifier].registerReplica(replica);
    end if
end for
```

Fig. 2. Replica advertisement related procedures

1) Manager Maintenance Protocol: When a manager is notified the loss of another manager controlling the same replicated data, firstly, it determines the identifier of the primary manager that will lead the reparation. This is a symmetric election process that exploits the list of alive peers maintained by each manager.

Secondly, it marks the lost manager as unavailable and adds the corresponding manager repair request descriptor within the local list of outstanding repair requests.

Thirdly, the manager that leads this reparation, allocates a local identifier for a new manager, computes the key k_{new} that identifies this new manager, then routes a request to deploy a manager with key k_{new} towards the overlay network of storage nodes. This request contains the list of existing replicas and managers. Upon its installation, the new manager notifies its arrival to existing managers controlling the same replicated object.

Upon the notification of the arrival of the new manager in replacement of a failed one, each manager removes the corresponding repair request from the list of outstanding requests to repair, then adds the new manager within the list of alive managers.

Finally, once the reparation is terminated, one re-examines the list of outstanding requests and re-processes any outstanding request for which the lost manager was the leader.

2) Data Replica Maintenance Protocol: When a manager is notified the loss of a data replica, it first adds the corresponding replica repair request within the list of outstanding repair requests. Then, this manager checks whether it is the leader for this replica loss?

The manager responsible of the lost replica allocates a new local replica identifier, computes a new replica key, then routes the request to create a replica corresponding to this key towards the overlay network of storage. To create a new

replica a node, one can simply get a copy of an existing replica. Once the new replica is created, *Pws* notifies its location to each manager in charge of that replicated data.

Upon the reception of the new replica location, each manager updates its list of replica locations, then removes the corresponding request to repair.

V. Analytical Evaluation

The objective is to compare the cost incurred by *Pws* with the one incurred by a basic leaf set-based replica maintenance system (BLS) to enforce replica availability in presence of node churn. We consider two metrics: the number of messages exchanged over the network and the network bandwidth consumed to enforce data availability. We consider a storage overlay that has already reached its equilibrium. We restrict the problem to the one of enforcing the availability of a set of already replicated objects. That is, none request to replicate a new object is issued any more.

The number of watchers enforced by *Pws* for each storing node is equal to the replication degree for both *Pws* and the leaf-set based replica maintenance system. Consequently, *Pws* maintains as many object replicas as replication managers per replicated object. Furthermore, we assume that *Pws* spreads out replicas and managers of each replicated object over distinct storing nodes.

Let consider the following notations: ω, the replication degree; γ the average total number of replicas (and hence of managers) stored at each node; μ, the average size of each object replica; and η the size of meta data maintained by a manager. Finally, let $\lambda = max(\eta, \tau_1, \tau_2, \phi_1, \phi_2, \phi_3, \phi_4, \phi_5)$, where: τ_1 is the size of the descriptor of a watcher; τ_2, the size of the descriptor of a replica as perceived by a watcher; ϕ_1, the size of the response returned by a watcher on the reception of new membership list; ϕ_2, the size of the response returned by a watcher on the reception of an advertisement; ϕ_3, the size of the request to participate to notifications of replica losses; ϕ_4, the size of the response to a request to participate to notifications; ϕ_5, the size of each replica loss notification. Note that these parameters are implementation-dependent.

A. Overhead on New Node Arrival

a) Pws Cost: When a new node joins the overlay network, this arrival impacts the watch set membership of the ω closest neighbors of the new node. Hence, in the worse case *Pws* runs a watch set maintenance procedure on each of the ω closest neighbors of the new node. For each concerned neighbor, *Pws* constitutes a watch set membership, then this new watch set to its members. This amounts to ω update requests (each request carries the new watch set membership), plus the same number of responses. In addition, on each watch set update, the watched node advertises its stored contents to its new watchers. With our assumption, at worse, there is one new watcher for each of the ω closest neighbors.

If $t4join_{pws}$ (resp. $bc4join_{pws}$) denotes the number of network messages (resp. number of bytes) transmitted over the network by *Pws* on each node arrival to enforce data availability, the following equations hold.

$$t4join_{pws} \leq 2\omega(\omega + 1) \quad (1)$$

$$bc4join_{pws} \leq \lambda(\omega^3 + \omega^2 + \omega + \gamma) \quad (2)$$

b) BLS Cost: With a basic leaf set-based replica maintenance, whenever a new node joins the overlay, it cooperates with its ω closest neighbors to determine the objects to replicate. Firstly, each of the ω closest neighbors of the new node computes the list of object identifiers that the new node should replicate, then sends this list to the new node. Each list contains in average γ/ω object identifiers. Later on, the new node will request each of its neighbors to send it each object that it should replicate. Once the new node has complete replica migrations, it informs each of its neighbors. If $t4join_{bls}$ (resp. $bc4join_{bls}$) denotes the traffic (resp. bandwidth consumed) on a node arrival by a leaf set-based replica maintenance system, the following equations hold:

$$t4joinbls \geq \omega + \gamma \quad (3)$$

$$bc4join_{bls} \geq \gamma\mu \quad (4)$$

c) Comparison: From [1] and [3], we observe that on node arrival, *Pws* will generate less traffic BLS provided that $\gamma > 2\omega^2 + \omega$. With respect to network bandwidth consumption, let r_j denote the ratio $bc4join_{pws}/bc4join_{bls}$; from [2] and [4], we derive $r_j \leq \lambda\mu^{-1}[1 + (\omega^3 + \omega^2 + \omega)\gamma^{-1}]$. This ratio indicates that the larger μ compared to λ, the better *Pws* compared to BLS. To illustrate, consider a prototype implementation that enforces $\lambda \leq 10^3$ and let $\omega = 4$ and $\gamma = 50$. If μ is equal to 10^6, then on network arrival BLS consumes 373 times more network bandwidth than *Pws*. It is worth noting that in [4] we consider only the bandwidth due to replica migrations.

B. Overhead on Node Departure

d) Pws Cost: When a node departs from the underlying network overlay, the underlying network management layer notifies this departure to nodes in its neighborhood, and hence to each member of its watch set. Upon a departure, in addition of the overhead (equivalent to the one incurred in case of a new node arrival) due to the watch set maintenance procedure, *Pws* incurs the following additional costs:

- Inter watchers cooperation. Watch set members cooperate with one another to determine which one is responsible to notify object losses to their suitable managers. This consists on $\omega - 1$ requests issued by secondary watchers and as much responses from the primary watcher (for more detail on inter watchers cooperation protocol, refer to Section IV-D).
- Loss notifications. For each entity stored at the departed node, one loss notification is sent to its manager. In average, there are 2γ loss notifications sent over the network.

 ISBN: 978-1-61208-280-6

- Loss repairs. To repair a replica (resp. replication manager) loss, its manager issues a request to create a new replica (resp. replication manager) at a random storing node. In average, on each node departure, there are γ requests to recreate replicas and γ requests to recreate replication managers. Note that each request to create a replica contains a means to help locate existing replicas, while a manager creation request carries the whole management state.
- Replica and manager recreations. To recreate a new replica, *Pws* requires 2 messages: one request to an existing replica to retrieve its state and a response containing the replicated state. Note that, to create a new manager, there is no need of network communication. Also, each newly created replica or manager is notified to the group of managers of its replicated object.

Overall, if $t4depart_{pws}$ (resp. $bc4depart_{pws}$) denotes the number of messages (resp. bytes) transmitted over the network by *Pws* on each node departure to enforce data availability, the following equations hold.

$$t4depart_{pws} \leq 2\omega^2 + 4\omega + 8\gamma \quad (5)$$

$$bc4depart_{pws} \leq \lambda[\omega^3 + \omega^2 + 2\omega + 8\gamma] + \mu\gamma \quad (6)$$

e) BLS Cost: With a basic leaf set-based replica maintenance system, whenever a node departs from the overlay network, a replica maintenance procedure is run by its ω closest neighbors in order to recreate replicas that are lost due to this departure. Let m designate one of the ω closest neighbors of the departed node. Firstly, the system computes the list of objects stored by the departed node for which m is (currently) the root. Secondly, for each such object, the system peaks one of the ω closest neighbors of m that doesn't yet replicates this object and sends it a request to create a replica.

On average, each m requests the recreation of $\gamma/\omega)$ replicas that were stored at the departed node. To recreate each replica, at least one request of the object state and one response containing the requested state are requires. Hence, if $t4depart_{bls}$ (resp. $bc4depart_{bls}$) denotes the number of messages (resp. bytes) transmitted over the network by a leaf set-based replica maintenance system, the following equations hold:

$$t4depart_{bls} \geq \omega + 2\gamma \quad (7)$$

$$bc4depart_{bls} \geq \gamma\mu \quad (8)$$

f) Comparison: On node departure, *Pws* incurs more traffic and consumes more network bandwidth to recreate managers and data replicas that are lost. From [6] and [8] it comes that if we consider large size objects (e.g., $\mu \geq 10^8$) both systems incurs comparable overhead on node departure. To see why, one could observe that as the average size of stored contents, the number of objects per node tends to diminish and so is the replication degree. We anticipate that, for large size objects both $\lambda\gamma$ and to $\lambda\omega^3$ are negligible compared to μ.

VI. Conclusion

We presented *Pws*, a peer-to-pee watching system that constitutes a replica maintenance substrate for DHT-based storage network. *Pws* that enforces a multiple publication keys replication approach, while avoiding an active tracking of storing nodes availability. We introduced the notion of node's watch set and detailed how to guarantee that watch set members remain in the neighborhood of the watched node in presence of node churn. We also presented the cooperation between storing nodes and watchers on the one hand, and between watchers and mangers on the other hand. The analytical evaluation of *Pws* overhead confirms that this system is an interesting alternative to maintain data availability for peer-to-peer storage utilities that destined to serve large-size objects.

References

[1] K. Kyungbaek and P. Daeyeon, "Reducing data replication overhead in dht based peer-to-peer system," in *Proceedings of the 2006 International Conference on High Performance Computing and Communications*, vol. 4208. LNCS, 2006, pp. 915 – 924.

[2] R. Bhagwan, K. Tati, Y.-C. Cheng, S. Savage, and G. M. Voelker, "Total recall: system support for automated availability management," in *Proceedings of the 1st conference on Symposium on Networked Systems Design and Implementation - Volume 1*. Berkeley, CA, USA: USENIX Association, 2004, pp. 337–350.

[3] K. Tati and G. M. Voelker, "On object maintenance in peer-to-peer systems," in *Proceedings of the 5th International Workshop on Peer-to-Peer Systems*, 2006.

[4] I. Stoica, R. Morris, D. Karger, M. F. Kaashoek, and H. Balakrishnan, "Chord: A scalable peer-to-peer lookup service for internet applications," in *Proceedings of the 2001 conference on Applications, technologies, architectures, and protocols for computer communications*, ser. SIGCOMM '01. New York, NY, USA: ACM, 2001, pp. 149–160.

[5] B. Y. Zhao, L. Huang, J. Stribling, S. C. Rhea, A. D. Joseph, and J. D. Kubiatowicz, "Tapestry: A resilient global-scale overlay for service deployment," *IEEE Journal on Selected Areas in Communications*, vol. 22, pp. 41–53, 2004.

[6] A. I. T. Rowstron and P. Druschel, "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems," in *Proceedings of the IFIP/ACM International Conference on Distributed Systems Platforms Heidelberg*, ser. Middleware '01. London, UK, UK: Springer-Verlag, 2001, pp. 329–350.

[7] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A scalable content-addressable network," *SIGCOMM Comput. Commun. Rev.*, vol. 31, no. 4, pp. 161–172, Aug. 2001.

[8] F. Dabek, B. Zhao, P. Druschel, J. Kubiatowicz, and I. Stoica, "Towards a common api for structured peer-to-peer overlays," *IPTPS03 International workshop on PeerToPeer Systems*, pp. 33–44, 2003.

[9] P. Druschel and A. Rowstron, "Past: A large-scale, persistent peer-to-peer storage utility," in *Hot Topics in Operating Systems, 2001. Proceedings of the Eighth Workshop on*. IEEE, 2001, pp. 75–80.

[10] A. Rowstron and P. Druschel, "Storage management and caching in past, a large-scale, persistent peer-to-peer storage utility," in *Proceedings of the 18th ACM Symposium on Operating Systems Principles*, 2001, pp. 188–201.

[11] S. Legtchenko, S. Monnet, P. Sens, and G. Muller, "Churn-resilient replication strategy for peer-to-peer distributed hash-tables," in *Proceedings of the 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems*, ser. SSS '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 485–499.

[12] A. Ghodsi, L. O. Alima, and S. Haridi, "Symmetric replication for structured peer-to-peer systems," in *Proceedings of The 3rd International Workshop on Databases, Information Systems and Peer-to-Peer Computing*, Trondheim, Norway, 2005, p. 12.

[13] M. Makpangou, "P2p based hosting system for scalable replicated databases," in *Proceedings of the 2009 EDBT/ICDT Workshops*, ser. EDBT/ICDT '09. New York, NY, USA: ACM, 2009, pp. 47–54.

 ISBN: 978-1-61208-280-6

Robust and Semi-automatic Electronic Health Record Dissemination Using the Devices Profile for Web Services

David Gregorczyk, Timm Bußhaus, Stefan Fischer
Institute of Telematics
University of Lübeck, Germany
{gregorczyk, busshaus, fischer}@itm.uni-luebeck.de

***Abstract*—Much conceptual work was done on medical device interoperability. Though many architectures, terminologies, and standards exist today, they have not achieved the desired acceptance yet, or provide ambiguous implementation directives. Recent research has shown that the Devices Profile for Web Services (DPWS) is suitable for device interconnectivity. Since DPWS was made to support a wide range of device types, it lacks concrete message process flows to enable added value in clinical environments. Hence, we first discuss existing approaches to interconnect devices in the operating room or intensive care. Afterwards, we introduce a protocol to transmit Electronic Health Records (EHRs) between hospital information systems and medical devices. Our evaluation shows that EHR distribution can be done almost automatically, while robustness is guaranteed against devices which will join a device ensemble both early and late, or against devices which will crash during intervention.**

***Keywords*-Web Services; DPWS; network protocols; e-health records;**

I. INTRODUCTION

Nowadays Operating Rooms (ORs) are equipped with numerous electronic appliances like navigation systems, operating microscopes, anesthesia machines, ventilators, and much more. While the amount of medical devices continuously increased over time, device interconnectivity has not been adapted in the same way [1]. The provision and consumption of medical information along the treatment path is either not existent or enabled by proprietary protocols, which are often very limited in terms of interoperability.

Generally, IEEE defines interoperability as the ability of two or more IT systems to exchange information and to utilize the information that has been exchanged [2]. By establishing interoperability in medical scenarios, right information would be provided at the right time, in the right amount, at the right location, and in the necessary quality [3]. As a consequence, caregivers would be relieved from workload and could benefit from decision support systems, resulting in higher patient safety. Some product solutions like OR.1 from Storz [4] or EndoALPHA from Olympus [5] already offer integrated medical device ensembles. Unfortunately, these complete systems prevent hospital operators from buying best-of-breed products. Beyond that, they do not cover plug-and-play-like device exchanges, which is an important feature for future integrated ORs.

So far, different international research and standardization activities pursue the establishment of medical device interoperability (see section II). This led to a lot of fundamental concepts. Some of the results were never used in practice, others miss concrete protocols for enhancing device collaboration or data exchange with Hospital Information Systems (HISs).

Recent work has revealed that a promising approach to interconnect medical devices are Service-oriented Architectures (SOAs), originally used in enterprise environments. SOA is a design paradigm for distributed IT systems. Functionality is provided in form of services. These services are connected by an abstract messaging backbone and offer the ability to be mutually consumed. By using a service directory mechanism, services are loosely coupled and can be dynamically exchanged at runtime. A more comprehensive introduction to SOA is given in [6]. By adapting a Service-oriented Device Architecture (SODA) [7], SOA principles were made applicable to (medical) devices.

Since SOA is just a conceptual model, technological specifications are required to implement a service-oriented environment. For clinical use, it has become apparent that the most common specification is the Web Service Technology. Invented in the early 2000, Web Services are based on the well-known Extensible Markup Language (XML) and the communication protocol SOAP. They further encompass a large set of extensions, called WS-*, meeting Web Services addressing, transactions, reliability, eventing, and further features. For more information please refer to [8].

Meanwhile, foundational interconnectivity concepts and technologies can be considered as sufficiently mature. As the next stage, investigations on elaborated protocols should be done to get medical devices start working together in a dynamic manner, generating added value. A frequently discussed use case in terms of medical device interconnectivity is the acquisition of Electronic Health Records (EHRs) to display patient demographics at the device side and re-use data for documentation affairs. An EHR is usually known as a systematic collection of digitalized patient information that should be sharable between different health care settings. We use EHR, patient data, and patient demographics synonymously.

In this paper, an automatic EHR dissemination process will be described to avoid the necessity to manually typing in

 ISBN: 978-1-61208-280-6

patient data at every single device involved in a surgery. Section II gives a survey of prior work related to medical device interoperability and patient record acquisition. Section III describes prerequisites and assumptions on our approach. In section IV protocol and implementation details will be explained. The protocol is evaluated in section V. Section VI concludes our work, enriched with some impressions on future perspectives.

II. Related Work

Many research projects and standardization efforts were carried out on medical device interoperability. Interesting work comes from Ibach et al. [9], Mauro et al. [7] and Pöhlsen et al. [10]. Ibach et al. introduce a SOA-based connectivity model including a dynamic service discovery mechanism and risk analysis. Mauro et al. modified the SOA paradigm to SODA (mentioned in Section I). Beyond SOA, their model contains a legacy wrapper pattern, a dynamic adapter pattern, and an auto-publishing pattern. Pöhlsen et al. have designed mechanisms and protocols for data security, reliability optimization, and discovery over subnet-boundaries. All approaches have in common that they depict Web Services as a middleware solution. Especially the Devices Profile for Web Services (DPWS) [11] plays a major role, because it comprises decentralized service discovery and publish/subscribe capabilities.

Further important work is done by Goldman et al. as part of the Medical Device "Plug-and-Play" Interoperability Program (MD PnP) [12]. They created the Integrated Clinical Environment (ICE) standard, defining functional elements for Point-of-care (PoC) related IT systems, especially focusing on communication of patient data, and on equipment command and control [13]. Though the ICE standard gives sophisticated information on conceptual system design, no concrete implementation details have been defined.

Besides, some standards deal with medical device interoperability. They are ISO/IEEE 11073, Health Level 7 (HL7) and Digital Imaging and Communications in Medicine (DICOM). While ISO/IEEE 11073 is specifically designed for device communication issues, the latter standards focus on data exchange between different clinical departments.

ISO/IEEE 11073 is a standards family separated into series 11073-1xxxx to 11073-7xxxx, of which the first three are the most important ones. ISO/IEEE 11073-1xxxx defines fundamentals for all subsequent parts, containing language elements, semantics, and an object-oriented Domain Information Model (DIM). The second part describes message exchange patterns between medical devices referring to the upper application layers of the ISO/OSI model. Physical interfaces are described as part of the 3xxxx serie. Today they are based on wired and wireless communication techniques (since infrared has never been accepted [14]).

HL7 is both a name for a not-for-profit, ANSI-accredited organization and a set of standards. It provides frameworks for integration and exchange of electronic health information between different vendors. Currently, two major versions exist: HL7v2 and HL7v3. Based on CSV-like formatted text data, HL7v2 is a pragmatic approach for message exchange, whereas version 3 uses XML and defines a comprehensive semantic Reference Information Model (RIM) of clinical processes.

DICOM is an open standard, preferably founded for the management of image data. It is typically used by radiology imaging systems, and supports encoding of one and two dimensional signal curves, and even video data. On top of that, it is possible to create work-lists and diagnosis-reports containing OR management data. Because the DICOM specification comprises more than 4000 pages, no end system supports the whole standard. Instead, DICOM conformance statements are used to confirm a certain set of functions.

In this context, another prominent initiative is Integrating the Healthcare Environment (IHE). Rather than specifying new standards, IHE is a group of health-care professionals and industry members harmonizing given standards and defining clinical processes on top of them.

All previous mentioned standards define conceptual workflows or data formats in terms of EHR dissemination. Unfortunately, most of them consider HISs, only. Otherwise, there are no implementation directives defined to technically accomplish data distribution in a plug-and-play like fashion. In this paper we propose to utilize concepts of the IHE Patient Demographics Query (PDQ) specification in connection with DPWS technology to give a concrete process flow for EHR distribution.

III. Prerequisites & Assumptions

Starting to think of data distribution between (medical) devices seems to be a simple task, but gets very complex the more details come into play. Therefore, this section makes some prerequisites and assumptions on work that would exceed the scope of this paper.

A. Devices Profile for Web Services

DPWS is defined as a set of Web Service standards tailored to be run on constrained devices. Some useful enhancements like device discovery over subnet-boundaries [10] and dual-channel transmission [15] complements this profile. Our feasibility study [16] has shown that DPWS is suitable for the clinical environment. Hence, it forms the communication framework for our concerns.

B. Authentication

Security plays a major role in integrated clinical environments, where data exchange is made among devices of different vendors and HISs. Enterprise security concepts are widely adopted within the hospital management, but not in the scope of medical device interaction. Beyond confidentiality and availability, dealing with EHRs requires two important security aspects: data integrity and accountability. It helps preserving patient safety and guarantees information usability in court cases. Integrity and accountability can be established by using authentication and non-repudiation mechanisms, enabled through digital signatures and data logging. WS-Security [17]

 ISBN: 978-1-61208-280-6

```
1<patient classCode="PAT">
2 <id root="1.2.840.114350.1.13.99998.8734" extension
      ="34827R534"/>
3 <statusCode code="active"/>
4 <patientPerson>
5  <name>
6   <given>Jim</given>
7   <family>Jones</family>
8  </name>
9  <telecom value="tel:+1-795-555-4745" use="HP"/>
10 <administrativeGenderCode code="M"/>
11 <birthTime value="19630713"/>
12 <addr>
13  <streetAddressLine>8734 Blue Ocean Street</
        streetAddressLine>
14  <city>Other City</city>
15  <state>IL</state>
16 </addr>
17 <!-- ... -->
18 </patientPerson>
19 <!-- ... -->
20</patient>
```

Listing 1. Sample HL7v3 EHR record [19] which is also part of the PDQ specification.

provides directives to handle digital signatures. Pöhlsen et al. [18] made a concept for distributed access control of medical devices including integrity and accountability. We assume that every exchanged message can be tested on these security parameters.

C. Context acquisition

Before EHRs can be safely transmitted, it is indispensable to ensure a common device communication context. This context helps grouping devices together such that they know each other and the subject they will be applied to. Unfortunately, it is not sufficient to simply argue that, for example, a sub-network provides a common device context. This consideration is obsolete because of wireless technology and even sub-networks which are spanned over more than one OR. Thus, it is mandatory to create infrastructure-independent device ensembles.

Context acquisition is done by designating a shared unique identifier either manually or automatically for a group of objects. Regarding to IEC 80001 [20], applying the context manually could be established by IT network risk managers the first time a device is placed in an OR. This works well for non-mobile units, but is not applicable to mobile devices. Gaining a context automatically by means of computer-supported localization techniques is a complete additional research area. Presently, we know no adequate way of automatic localization. Hence, for our protocol we assume that every device has already acquired contextual information.

D. Semantics

One condition to produce inter-operable IT systems is standardization of data formats and semantics. Only if every parameter is strictly regulated, devices of different types and vendors can work together. By using DPWS, data is XML-serialized by default, and can be structured and described with XML Schema. Meaningful data could be generated by applying semantic identifiers like they are defined in the HL7v3 RIM. The IHE initiative proposes to use a subset of HL7 terminology. An example data set is illustrated in Listing 1. As we can figure out, foundational elements are already included like name, gender and birth date. Beyond that, the inclusion of weight, size, medication information, and even Web Access to DICOM Persistent Objects (WADO) [21] could also be useful. The process of disseminating data is independent of any message payload, so we consider patient data to be given by a third party. Since XML offers dynamic language extensions, data is addable on demand.

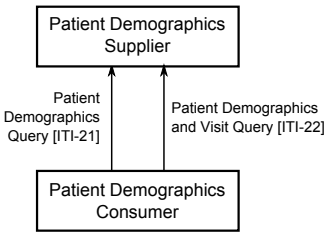


Fig. 1. IHE PDQ architecture as described in [22].

IV. EHR Dissemination Protocol

The acquisition of EHRs are twofold. First, what type of data should be provided, and second, which steps are necessary to enable data distribution. As mentioned above, the first aspect is out of scope. Regarding the second point, the IHE initiative has created an IT Infrastructure Technical Framework (ITI TF) [22], including architectures and transactions to obtain patient data. Fig. 1 shows the IHE proposed architecture to resolve EHRs. Patient data consumers send *Patient Demographics Queries* or *Patient Demographics and Visit Queries* to a patient supplier. The supplier in turn receives these requests, obtains data by means of proprietary interactions with third party systems, and returns them to the requesting consumer. This system works great if there is only one device and the caregiver has to confirm patient data once. As soon as two or more devices need data, a significant amount of additional work is generated: the caregiver has to confirm patient data at every single device. Furthermore, if a device crashes or is rebooted anyway, data has to be requested and confirmed again. It is likely that such systems will not be accepted by clinical staff.

A. Coarse-grained Procedure

Fig. 2 depicts the physical device infrastructure of an OR. Every connected device and IT system know their neighbors by means of WS-Discovery (WS-DD) [23], that is part of DPWS. Devices are grouped by using the context identifier as a scope parameter of a WS-DD Probe request. Therefore, context identifiers should be representable as a URI.

Typically, an OR Management System (ORM) builds a bridge to the HIS, using protocols like DICOM and HL7. In the following, an IT system that fetches EHRs, will be referred as a gateway unit. Since most ORMs speak protocols that will not suit to DPWS (even Web Services based on Basic Profile 1.1 differ in the SOAP version that DPWS prescribes),

 ISBN: 978-1-61208-280-6

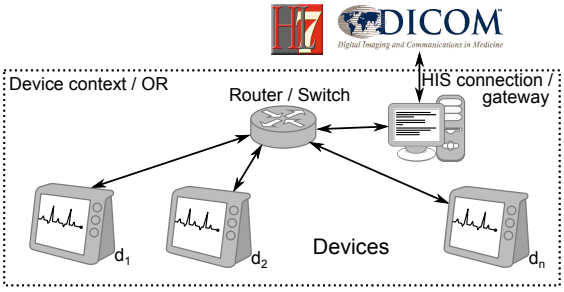


Fig. 2. Physical infrastructure of an OR. The HIS might be connected through a switch/router or—like it is illustrated here—connected through an OR Management System.

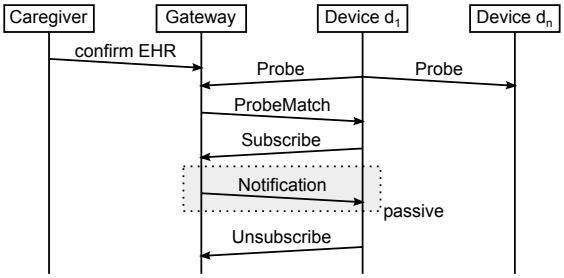


Fig. 3. Sequence diagram of the basic process flow. A notification is send once per intervention. To keep things clear, the process is illustrated for d_1 only.

a certain converter unit is required to translate messages. Since our system is SOA-based, theoretically any device could play the converter role. However, in most cases the converter will be deployed on the gateway unit or on a separated computer.

The process flow comprises four stages:

1) EHRs (please note the plural) of different patients are requested and will be stored on the gateway unit. Since DPWS is SOA-based, it is possible to deploy the gateway on the ORM or on any medical device.
2) As soon as EHRs are available, they are transferred to the converter unit. Usually, ORMs do not actively publish their data, so that the converter unit has to periodically pull them.
3) A certain patient record must be confirmed. This step is required due to regulatory affairs. It implicates a human actor and a human-machine-interface either on the gateway unit or on any other medical device. The caregiver authenticates to the system and confirms a record out of a list. This list could be filtered, e.g., by date, by using bar codes or near field communication, or by means of a mobile unit that provides patient information. Finally, in every case the caregiver has to confirm an EHR due to legal reasons.
4) Data is distributed by using a passive and active request sequence. This stage is described in detail in subsection IV-B.

B. Dissemination Protocol

Electronic, computer-supported systems or devices always suffer from failures and will seldom be used as intended. Because of that, it is very important to design failure resistant systems. For our concerns, to compensate connection losses and reduce workload overhead, we introduce a passive and an active EHR retrieval using WS-DD and WS-Eventing (WS-E) [24]. Passive means that data is distributed via publish/subscribe. Any medical device, which is interested in patient data, subscribes to a gateway unit and passively waits for incoming messages. Active means that any medical device is getting patient data by directly sending a request to the gateway unit.

A (desirable) process flow is given as depicted in Fig. 3. First, every device d_i sends a WS-DD Probe request with a proper context id. To achieve a more precise result set, an EHR supplier type could be defined. Due to the fact that WS-DD scopes and types can be freely selected, they have to be standardized to ensure interoperability. After a suitable gateway g was found, devices d_i have to send a WS-E Subscribe message. The delivery mode [25] and filter dialect [26] are prescribed by DPWS. Similarly to scopes and types, filter URIs have to be standardized. The parameter `Expires` must be set to any realistic value, and the underlying subscription should be renewed accordingly. Any subscription runs until the corresponding software is logged out from the system or shut down. As soon as an EHR is available, it will be published via WS-E Notification messages.

The aforementioned process flow discusses an optimal behavior, but sometimes devices will not be available, when the gateway is ready to send EHR notifications. Beyond, it is even desirable to confirm patient data at the device side, not exclusively at the gateway. To overcome these issues, another step is required: the active mode. As soon as a device has subscribed to the gateway unit, it asks for any existing EHR. The gateway responds either with a single confirmed record or, if no patient data was confirmed yet, with a list of records. In the latter case patient records can be confirmed at the device side and then be sent back to the gateway. Since the gateway computer is the central instance to manage EHRs, the device on which the patient was confirmed, has to wait for the WS-E Notification (see Fig. 5) instead of using the recently confirmed record. Any WS-Addressing action identifiers for these requests, or any mechanisms to confirm patient records, are out of scope and have to be defined by a standardization committee.

The active and passive process flow is shown in Fig. 4. Due to communication delays, it could be possible that a device actively requests patient data and receives an additional notification. Usually, EHRs contain an unambiguous identifier representing the complete patient's hospital stay. If this identifier is missing, an alternative approach is necessary to guarantee EHR uniqueness. Therefore, EHRs should be signed with a timestamp/clock/node based UUID [27], designated as u_{origin}. If a device resolves patient information the first time, it persists this UUID as u_{last}. When another EHR is received

 ISBN: 978-1-61208-280-6

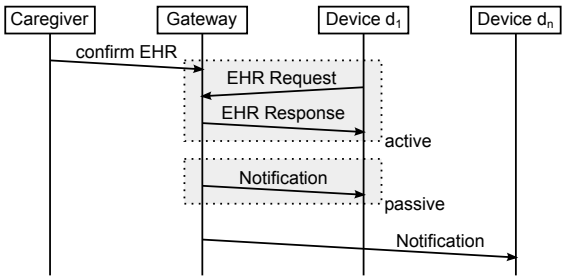


Fig. 4. Sequence diagram of the enhanced process flow. Here the patient record is confirmed in advance. Devices d_i fetch the data actively and passively. It should be noticed that EHRs might be confirmed at any point in time. To keep things clear, the process is illustrated for d_1 only. Any probes and subscriptions were already performed.

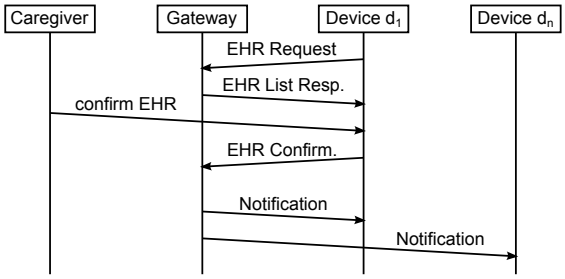


Fig. 5. This diagram shows the process of confirming an EHR at the device side. First, a requesting device receives a list instead of a single confirmed record. Then it visualizes the list and waits for data confirmation. Afterwards, the confirmation acknowledgment is send to the gateway followed by a WS-E Notification—only the notification data is relevant to the device.

or requested, $u_{origin} = u_{last}$ indicates that patient data had not changed, while $u_{origin} \neq u_{last}$ reveals that a new patient was allocated.

On the other side, if a gateway computer crashes, a WS-DD Hello message indicates that devices have to re-subscribe to the gateway.

V. Evaluation

The goal of our work was to create an automatic and robust patient data dissemination protocol. To reduce workload, data is distributed almost in an automatic manner using WS-DD and WS-E. By sending EHR requests and waiting for WS-E Notification messages, devices are able to automatically resolve current patient data independent of system crashes and early or late group joining. We do not consider the case when a device discovers two or more EHR suppliers during a Probe request. Due to SOA's lose coupling directive, every supplier should provide the same patient information consistent with the current HIS/ORM data. A device could benefit from several data sources by performing plausibility checks. Since plausibility checks are not part of our communication protocol, they were not considered any further.

A. Experimental setup

To evaluate the performance and feasibility of our system, a prototypical implementation has been set up on a

PortType WS-Eventing	PortType WS-Discovery	PortType EHR Service
Operations: - Subscribe - Renew - GetStatus - SubscriptionEnd - Notification	Operations: - Hello - Bye - Probe - ProbeMatch - Resolve - Resolve Match	Operation: - GetEhr Subscription: - EhrConfirmation

Fig. 6. Interface definition required to disseminate patient data. Most functionality is derived from well-known standards, which are also part of DPWS. WS-E operations are defined in [24] whereas WS-DD operations are defined in [23]. Any EHR supplier has to implement at least a *GetEhr* operation for active mode and an *EhrConfirmation* event source for passive mode.

Microsoft Windows 7 64-bit based Java Virtual Machine using the Web Services for Devices Java Multi Edition Stack (WS4D-JMEDS), version 2.0beta8 [28]. Fig. 6 illustrates the WSDL-based service interface a gateway has to realize. Two kinds of measurements are taken. First, the amount of messages exchanged to disseminate EHRs. Second, the elapsed time until every communication participant has received patient records when using 1, 2, 4, 8, 16, 32, and 64 consumers. Since our approach is a novel consideration to automatically enable patient data on multiple medical devices, it is difficult to compare with other systems. A general performance gain is fundamentally given through the fact that just a single EHR confirmation is required instead of N confirmations, where N is the amount of devices interested in any patient data. The underlying setup uses no authentication yet.

The experimental setup comprises of a single PC, equipped with a 2.4 GHz quad core CPU and 16 GB of RAM. All devices, including the gateway, run on one machine within different Java processes. In the following, they are called virtual devices. Payload data is taken from the IHE PDQ [19]. It is about 10 KB of data. To take time measurements, a monitoring application is connected to the gateway and consuming devices d_1 to d_n. This application is responsible to collect event triggers by using a primitive Java Remote Method Invocation (RMI) application. Since virtual devices run on a single PC, no network traffic is generated. Hence, in real-world scenarios additional transmitting time has to be expected. Furthermore, Java cross-optimizes shared objects, which will lead to additional performance gains.

To illustrate feasibility of the approach, following test cases are considered:

1) The gateway is started and the EHR is confirmed. Hereafter, devices d_1 to d_n are turned on.
2) The gateway is started. When booted, devices d_1 to d_n are turned on. Then, the EHR is confirmed.
3) The gateway is not available. Devices d_1 to $d_{\frac{n}{2}}$ are turned on. When done, the gateway is started and the EHR is confirmed. Afterwards, devices $d_{\frac{n}{2}+1}$ to d_n are turned on.
4) When patient data is distributed, a single device is synthetically shut down and turned on again to simulate

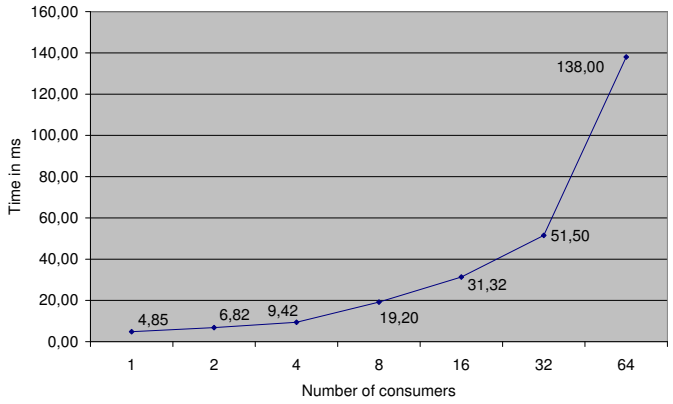


Fig. 7. Different measurements of time span between confirming an EHR and disseminating it to interested sinks.

a crashed PC. This is done by sending and even omitting a WS-DD Bye message.

5) When patient data is distributed, a gateway crash is simulated. This is also done by sending and omitting a Bye message.

B. Results

Regarding the foregoing test cases, point one and two work as expected. In the first case, data is retrieved using the active mode (operation *GetEhr*). In the second case, data is retrieved using the passive mode (event source *EhrConfirmation*). Test case three forces d_1 to $d_{\frac{n}{2}}$ to wait in passive mode, while $d_{\frac{n}{2}+1}$ to d_n can obtain their data using the active mode. The fourth scenario is twofold. If the single device d_i is shut down sending a Bye message, it can properly be removed from the gateway's subscription list. In the other case, the gateway does not know if d_i is out of order. To avoid maintaining obsolete subscriptions, client devices are encouraged to send heartbeats in form of WS-E Renew messages. After d_i is restarted, data is retrieved using the active mode. To detect a patient exchange, d_i can compare the unique EHR identifier described in section IV-B. In the last case an EHR supplier crash is simulated. If a Bye message is sent, every device clears their subscription and renew them as soon as the gateway enters the network again. If no Bye message is sent, client devices will not notice any changes. The gateway persists subscriptions and send WS-E SubscriptionEnd messages after restart. Client devices receive these messages and quit the outdated subscription to subsequently create a new one.

Fig. 7 shows the efficiency measurements in terms of the time span between confirming an EHR record and disseminating it to interested parties. Hundred measurements were taken for 1, 2, 4, 8, 16, 32, and 64 consumers. Having a rough overview of Fig. 7, doubling the amount of consumers causes a doubled time consumption. However, due to the fact that four cores process the gateway and consumers, time is not exactly increasing proportionally with the amount of consumers. In the end, there is no unreasonable time delay when distributing data up to 64 devices. Unfortunately, the monitoring instance had noticed data loss in some measurement runs. The reason for that behavior is not conclusively clarified. May be there is a software bug occurred in JMEDS, or there is an operating system issue when processing loads of TCP connections pointing to the localhost.

TABLE I
TRANSMITTED MESSAGES PER NUMBER OF CONSUMERS.

Number of consumers	Transmitted messages
1	13
2	26
4	52
8	104
16	208
32	416
64	832

The amount of transmitted messages required to disseminate EHRs is based on a theoretical approximation. Hereby, sending a SOAP message counts as sending one message. Sending a multicast message conforms to sending K messages, where K is the number of multicast channel subscriptions. The minimum amount of messages M is given by the following formula:

$$M = 13x + xp + \lfloor \frac{r}{t} \rfloor x$$

The variable $x \in \mathbb{N}$ is the number of consumers, $r \in \mathbb{R}_+$ is the time until a WS-E Renew message has to be sent, $t \in \mathbb{R}_+$ corresponds to the time until the gateway is shut down, and $p \in \mathbb{N}_0$ is the amount of patient exchanges occured during system runtime. This formula does not cover any failures and any devices joining the network late in time. It also does not cover the existence of further devices providing any services, which would increase the amount of WS-DD Hello and Bye messages. Therefore, the formula defines a lower bound for the amount of transmitted messages. The term $13x$ describes the amount of messages initially transmitted, comprising of discovering devices, subscribing to the gateway, quitting subscriptions and network participations, and retrieving patient data the first time. The term xp covers notifications sent by the gateway due to a patient exchange. The last term represents the number of messages transmitted to renew subscriptions. In our test scenario, it is: $t < r$ and $p = 0$. Hence, the amount of messages is simply based on $M = 13x$. TABLE I illustrates different message counts. In comparison to simple get requests like they are done using PDQs, the amount of messages is very large. But it is reasonable regarding to the benefit of reduced workload and a discovery process to dynamically plug in any devices.

VI. CONCLUSION & FUTURE WORK

In this paper, we have created an EHR dissemination protocol suited to dynamic, interconnected ORs meeting robustness against system crashes and early or late group joining. By using a passive and active mode, every device is synchronized

 ISBN: 978-1-61208-280-6

with the latest EHR. In comparison to existing systems no additional, human-involved configuration or confirmation is required, which reduces or at least does not increase the caregiver's workload. Due to the best-effort behavior of Ethernet, this protocol does not address transmission failures. Hence, it is not robust against physical interferences.

Further research has to be done both on mechanisms to handle documentation affairs on the basis of transmitted EHRs, and on localization techniques to (semi-)automatically compose device ensembles. Beyond that, optimizations regarding to WS-E filter techniques could be useful to reduce data load. Finally, the evaluation could be extended to real-world device setups with authentication enabled.

REFERENCES

[1] K. Lesh, S. Weininger, J. M. Goldman, B. Wilson, and G. Himes, "Medical Device Interoperability-Assessing the Environment," in *Proceedings of the 2007 Joint Workshop on High Confidence Medical Devices, Software, and Systems and Medical Device Plug-and-Play Interoperability (HCMDSSMDPnP)*. Cambridge, Massachusetts, USA: IEEE Computer Society, June 2007, pp. 3–12.

[2] "IEEE Standard Computer Dictionary. A Compilation of IEEE Standard Computer Glossaries," *IEEE Computer Society Press*, p. 1, January 1991.

[3] A. Schweiger, A. Sunyaev, J. M. Leimeister, and H. Krcmar, "Toward Seamless Healthcare with Software Agents," *Communications of the Association for Information Systems (CAIS*, pp. 692–709, 2007.

[4] Karl Storz, "Kartl Storz OR1," 2012/10/19, Version: 2013-05-13. [Online]. Available: http://www.karlstorz.com/cps/rde/xchg/SID-CF744F30-788546A4/karlstorz/hs.xsl/522.htm

[5] Olympus, "Olympus - EndoALPHA," Version: 2013-05-13. [Online]. Available: http://www.olympus-europa.com/endoalpha

[6] E. Thomas, *SOA Principles of Service Design*, 1st ed. Prentice Hall International, Jul. 2007.

[7] C. Mauro, A. Sunyaev, J. Leimeister, and H. Krcmar, "Standardized Device Services - A Design Pattern for Service Oriented Integration of Medical Devices," in *2010 43rd Hawaii International Conference on System Sciences (HICSS)*, Jan. 2010, pp. 1–10.

[8] C. Werner and S. Fischer, *Semantic Web Services: Concepts, Technologies, and Applications*. Springer, 2007, ch. 2, p. 25ff.

[9] B. Ibach, J. Benzko, and K. Radermacher, "OR-Integration based on SOA – Automatic detection of new Service Providers using DPWS," *International Journal of Computer Assisted Radiology and Surgery*, vol. 1, pp. 195–196, 2010.

[10] S. Pöhlsen, S. Schlichting, M. Strähle, F. Franz, and C. Werner, "A Concept for a Medical Device Plug-and-Play Architecture based on Web Services," *SIGBED Rev.*, vol. 6, no. 2, pp. 1–7, Jul. 2009, Version: 2013-05-13. [Online]. Available: http://doi.acm.org/10.1145/1859823.1859829

[11] E. Zeeb, G. Moritz, D. Timmermann, and F. Golatowski, "WS4D: toolkits for networked embedded systems based on the devices profile for web services," in *2010 39th International Conference on Parallel Processing Workshops (ICPPW)*, Sep. 2010, pp. 1–8.

[12] "MD PnP Program," Version: 2013-05-13. [Online]. Available: http://www.mdpnp.org

[13] J. M. Goldman, "Medical Devices and Medical Systems — Essential safety requirements for equipment comprising the patient-centric integrated clinical environment (ICE) — Part 1: General requirements and conceptual model," ASTM International, 2008, Version: 2013-05-13. [Online]. Available: http://www.mdpnp.org/uploads/F2761_completed_committee_draft.pdf

[14] M. Galarraga, L. Serrano, I. Martinez, and P. de Toledo, "Review of the ISO/IEEE 11073 - PoCMDC standard for medical device interoperability and its applicability in home and ambulatory telemonitoring scenarios," Electrical and Electronic Engineering Dept. Public University of Navarra (UPNA). Pamplona. Aragon Institute of Engineering Research (I3A). University of Zaragoza (UZ). Zaragoza. Biomedical Engineering and Telemedicine Group (GBT). Polytechnic University of Madrid (UPM). Madrid., Spain, Tech. Rep., 2006.

[15] S. Pöhlsen, W. Schöch, and S. Schlichting, "A Protocol for Dual Channel Transmission in Service-Oriented Medical Device Architectures based on Web Services," in *3rd Joint Workshop on High Confidence Medical Devices, Software, and Systems & Medical Device Plug-and-Play Interoperability*, 2011.

[16] D. Gregorczyk, T. Bußhaus, and S. Fischer, "A Proof of Concept for Medical Device Integration using Web Services." IEEE, Mar. 2012, pp. 1–6, Version: 2013-05-13. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6198124

[17] A. Nadalin, C. Kaler, and R. M. P. Hallam-Baker, "Web Services Security: SOAP Message Security 1.1 (WS-Security 2004)," OASIS Open, Tech. Rep., February 2006.

[18] S. Pöhlsen, S. Schlichting, and C. Werner, "Praktische Umsetzung einer verteilten Zugriffskontrolle für Webservices anhand ihrer Eigenschaften," *PIK – Praxis der Informationsverarbeitung und Kommunikation*, vol. 33, no. 2, pp. 140–146, Jun. 2010, Version: 2013-05-13. [Online]. Available: http://dx.doi.org/10.1515/piko.2010.021

[19] Jan. 2013, Version: 2013-05-13. [Online]. Available: ftp://ftp.ihe.net/TF_Implementation_Material/ITI/examples/PDQV3/02_PDQQuery1Response.xml

[20] S. Eagles, "An Introduction to IEC 80001: Aiming for Patient Safety in the Networked Healthcare Environment," *IT Horizons*, vol. 4, pp. 15–19, 2008.

[21] *DICOM Standards Commitee, 2010b. Supplement 148: Web Access to DICOM persistent Objects by means of Web Services: Extension of the Retrieve Service (WADO Web Service)*, National Electrical Manufacturers Association (NEMA) Std., 2010, Version: 2013-05-13. [Online]. Available: ftp://medical.nema.org/medical/dicom/Supps/sup148_17.pdf

[22] "IHE IT Infrastructure (ITI) Technical Framework – Volume 1 (ITI TF-1) Integration Profiles," IHE International, Inc., Tech. Rep., 2012.

[23] V. Modi and D. Kemp, "Web Services Dynamic Discovery (WS-Discovery) Version 1.1," OASIS Open, Tech. Rep., Jul. 2009.

[24] D. Davis, A. Malhotra, K. Warr, and W. Chou, "Web Services Eventing (WS-Eventing)," W3C, Tech. Rep., 2011.

[25] "WS-Eventing Delivery Mode," http://schemas.xmlsoap.org/ws/2004/08/eventing/DeliveryModes/Push.

[26] "WS-Eventing Filter Dialect," http://docs.oasis-open.org/ws-dd/ns/dpws/2009/01/Action.

[27] P. Leach, M. Mealling, and R. Salz, "RFC 4122: A Universally Unique IDentifier (UUID) URN Namespace," IETF, Tech. Rep., Jul. 2005, Version: 2013-05-13. [Online]. Available: http://www.ietf.org/rfc/rfc4122.txt

[28] "JMEDS (Java Multi Edition DPWS Stack)," Version: 2013-05-13. [Online]. Available: http://sourceforge.net/projects/ws4d-javame/files/ws4d-jmeds-v2/framework/v2.0.0-beta8/

 ISBN: 978-1-61208-280-6

Semantic Web Services Adaptation and Composition Method

Hamid Mcheick
Department of Computer Science and Mathematics
University of Quebec at Chicoutimi (UQAC)
Chicoutimi, Canada
hamid_mcheick@uqac.ca

Amel Hannech
Department of Computer Science and Mathematics
University of Quebec at Chicoutimi (UQAC)
Chicoutimi, Canada
amel.hannech1@uqac.ca

***Abstract*—Web services adoption is a major advance in the development of interoperable information systems. In particular, the composition of services can meet the needs increasingly complex of user, by a combination of web services within a single business process. However, despite this widespread adoption of Web services, many obstacles prevent their reconciliation in the composition, or may occur within a BPEL process in a state change, the context for example. ASWSCC Method (Adaptation of Semantic Web Service Composition to Context) is an implementation of a theoretical model made in our an earlier work. It focuses on composition process adaptation to use context (preferences, user type and its environment as the device used, location, access mode and many others). This context and request service matching should be taken into account while composing new services. Our goal is to develop a model which ensures, on the one hand, web services matching during composition process by using domain ontology as lexical database WordNet, its purpose is to identify, classify and relate in different ways semantic content and lexical language. On the other hand, this model allows management and taking into account the context that makes composition process adaptable to different instances of use context, which may change during the same session. For this reason, we are interested to capture and manage the context and its impact on basic services and composition process at once. Changes can affect the context of web services during their executions and the need to adapt their dynamically becomes increasingly crucial. From here comes the need for a coherent solution to adapt web services context. We exploit the benefits of aspect weaving tool in this approach to inject aspects of web services to adapt them to change of context.**

Keywords-Context definition and management; adaptation; web services composition.

I. INTRODUCTION

Internet evolution and competition between firms were factors in web services explosion services. Indeed, web services may constitute speed and efficiency contribution for e-business. This notion of web service essentially means an application made available on Internet by a service provider and accessed by clients through standard internet protocols [1]. Their characteristics compared to other distributed computing technologies lie in the fact that they offer component model in weak coupling using Internet technology as infrastructure for communication. If application designer target is not achieved by invoking a simple web service elementary, designer must combine the functionality of a set of services.

This process is called web service composition. It specifies which services need to be invoked in what order and how to manage interactions between them, and exception conditions. Service-oriented architecture (SOA) enables integration of applications and resources flexibly, representing every application or resource as a service. In particular, SOA suffers from a number of limitations and weaknesses in the context of composition on demand; hence web services adaptation to context remains essential to better exploit services. We introduced a basic idea of our service composition model in [5]. This paper describes the steps, a case study and comparison with exiting methods details of our service semantic composition method. The case study is a proof-of-concept to illustrate our service composition method (ASWSCC).

This article is organized as follows. Section 2 describes related works. Section 3 shows our process of service composition based on context, in which we describe the process of service composition. The validation of this model, through a case study and a comparison, are given in Section 4. Aspect-Oriented computing is illustrated in Section 5. Section 6 compares service compositions approaches. A conclusion and future work are discussed in Section 7.

II. RELATED WORK

Web services adaptation to contextual changes can occur at several levels. To adapt a process composition with Business Process Execution Language (BPEL), several researchers act on composition and orchestration of services. In their approaches, Keidl et al. [2] present a context framework that facilitates context-aware Web services development and deployment. In their framework, context information is exchanged in the header of Simple Object Access Protocol (SOAP) messages. In their proposal, authors use pre-defined Web services with contextual information clients. This information will be used later to provide a service to custom behavior.

Processing context is provided by web services, context plug-ins, or context services. Messages exchanged before and after this operation are based on contextual information

 ISBN: 978-1-61208-280-6

and are essential for processing and automatic adaptation to changes in web services context without need to adapt manually web services. Once context information processing is made, it is to component Invocation Manager to invoke the web service properly. It has already been pre-set with a context similar to that returned by the Context Manger. In their approach, Keidl et al. includes a new component called Context Manager responsible for handling context information customer.

The assignment of this operation to a full-fledged component seems to us very useful, because it allows distributing adaptation operation load, and on the other hand makes the approach more flexible especially when web services are defined as activities of BPEL process. In other words and in our case we can place this component in the form of activities in the BPEL process concerned.

Other work that is akin to our research problem is that of Chaari et al. [3] who defined three techniques of adaptation to web services context: Web service internal adaptation, Web service external adaptation, and Web services polymorphic adaptation. We can say from these three adaptation techniques that the first is a web service manual adaptation; the second is a dynamic adaptation, while the third is an adaptation in services composition. Among the works that have addressed such as adaptation, we cite the work of Vukovic and Robinson [4] who presented a system architecture for building context-aware applications based on the notion of Web services dynamic composition. In their approach, any contextual change can lead to new reconstruction during the execution of services, resulting thus dynamic evolution of the application. They developed a Framework which uses planning shop2 [13], BPEL4WS [4][21] and BPWS4J [22] for composition, design and execution of the composite service, respectively. Otherwise, the context changes will be doing manually from a client interface. Using context values as input, their service composition engine generates the query composition services suitable to adapt their application to these changes in context.

We learned from this approach using a client interface for entering information on the context. Thus, we can be inspired to capture the context information for client and allow him to change its environment even during the execution of the composition process.

Recent work has focused on semantic description web services and ontologies are mainly used to model the semantic service representation. It helps to establish semantic relations between concepts of the domain under consideration. We also have to mention that the OWL-S [20] approach that uses the ontology OWL-S to extend UDDI with semantic description of Web services.

III. PROCESS OF ADAPTATION OF SEMANTIC WEB SERVICES COMPOSITION TO CONTEXT

In a previous work, we have built a prototype system for Semantic Web services composition, we have shown through this system for use benefit of behavioral descriptions of Web services (class, function, operations, input, output, context, etc.), the description of use context and their contributions to semantic Web services composition, as well as the interest of taking into account the alignment of ontologies and similarity measures between web services candidates concepts during composition process [5]. In this paper, we present our case study and experiment validates the proposed prototype.

We start first by giving definition and context modeling approach that will be taken into account during web services selection process and composition due to a user request.

A. Contexte Definition

Many works from computer field sensitive to context try to give a definition of context to establish a basis for the adaptation process, but they have not yet resulted in a definition that is both generic and pragmatic context, and more precisely parameters constituting context. The most common definition in the literature and accepted by most researchers is that proposed by Dey et al. [6]. The latter defines context as « any information that characterizes the situation of an entity. An entity is a person, place or object considered relevant with respect to the interaction between a user and an application, including the user and application themselves». The context describes user situation in terms of user profile, environment, terminal used, location, time, etc. In addition, contextual data are completely independent of applications and are not retrieved from a storage medium connected to the field of application. From a practical point of view, a context can be given as it cannot be supplied by the user when invoking web services. More concretely, we define context as the set of external parameters that can affect the behavior of web services involved in the process of basic composition and composition process itself by defining new web service candidates constituent other dial plans. These parameters have dynamic appearance allowing them to change during the runtime. For example, during an operation « buy online book » by a student, a contextual situation can be defined with the following parameters: (access type = «Student», type = device «Smartphone», location = «UQAC»). A new value of these parameters presents a new contextual situation that can change all selected web services, and thus, change the execution plan process composition of these services.

To define a contextual situation more precisely, and based on a previously defined pattern in a patient records management application [23]; we define a three-dimensional space where each dimension represents an axis of context that covers: user type «access type», device type, and location. A change in the value of these parameters defines a new contextual situation in which the composition process must adapt. To better explain our definition, a space E (x, y, z) is explained in Figure 1. The state E1 represents a contextual situation in which the user is a student located in the library of the university using his mobile phone to buy a book for his class PHP development. The state E2 represents a contextual situation in which the user is a student in a building still inside the university but outside the library using a desktop PC to buy a book for his course PHP. The state E3 represents a contextual situation where the user is a student not located within the university using a desktop PC trying to buy a book for a class for PHP development.

 ISBN: 978-1-61208-280-6

Each of these contextual parameters causes a change in the search result. Despite using the same query "buy book PHP development" the answer is not the same for these three users, and the difference in contextual situation. Some services are only available to students «access type = Student» and only using the university's computer network «location = UQAC Network». In addition the university offers online books that are only accessible for students.

The device modifies the display mode and interaction mode with user (cutting all the information into subgroups to the small terminal, a single window with the entire information on the desktop screen, a graphical display of multidimensional arrays on a PC, texts speech synthesis on a mobile phone and smart phones). This implies that services that correspond to the parameters of location and type of access may not be compatible with mobile phones for example, and then they will remain unavailable «hidden» to user.

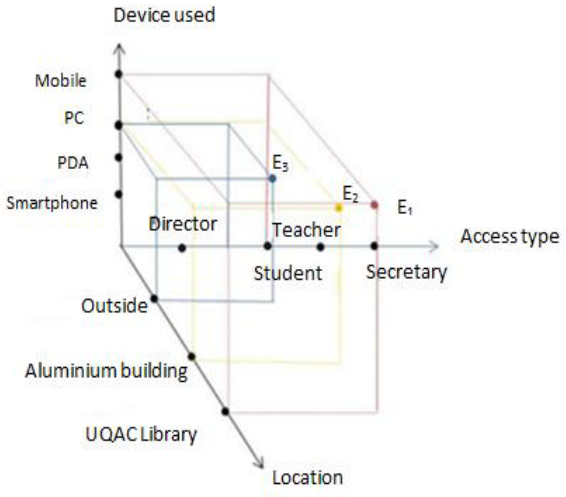


Figure 1. Representation of Context Categories

The device modifies the display mode and interaction mode with user (cutting all the information into subgroups to the small terminal, a single window with the entire information on the desktop screen, a graphical display of multidimensional arrays on a PC, texts speech synthesis on a mobile phone and smartphones). This implies that services that correspond to the parameters of location and type of access may not be compatible with mobile phones for example, and then they will remain unavailable «hidden» to user.

B. Management and Modeling Context

We store context using a set of pairs (attribute, value). For example, ContexteUtilisation (id_ContexteUtilisation = «context1», username = «Amel», attribute = «device type», value = «PDA»). (id_ContexteUtilisation = «context2», username = «Amel», attribute = «location», value = «UQAC»).

We define three aspects to model context (access type, device type, location). We also propose to store the context before its release in the selection process and web services composition, to keep track of historical values captured. This allows us to have a rich and reliable representation of data captured in form of ontology that represents the context as a set of entities describing their aspects.

C. Composition Process

Composition process starts by specifying the set of tasks to be performed that allows expressing composition purpose. This activity is triggered by a user request, it determines and organizes tasks to be performed by web services, each of which can perform one or more tasks in a dialing plan. This is done by decomposing user request represents a complex task into simpler and functional tasks, based on a knowledge base, this step is explained in out precedent word [5].

Planning is an essential step. Indeed, it determines web services that will participate in the composition. In addition, it determines web services ordering present in composition. Each dial plan is based on all sub queries such that each one contains its own semantic web services returned after discovery stage, and are used with others to build a plan or plans of dynamic composition, from the first sub query to the last.

Composition plan starts from the first sub query to the last, as we apply semantic matching algorithm [19] between the Output parameter of all semantic web services candidates from the first sub query with input parameter of all web services from the second sub query, and so, with the second sub query to the last.

Web services discovery that meet each sub task is based on semantic matching, it can search the descriptions of web services that have a semantic correspondence between the functional parameters defined in these descriptions (category, function, output, input, etc.) and those introduced into sub tasks. This is based on global domain ontology «WordNet» by comparing terms and gives an approximation between concepts [7]. The approximation is a degree of semantic similarity [19] to determine which concepts equivalents to query words which are considered acceptable by the system to generate an appropriate response.

So, the semantic matching algorithm [19] is based on the similarity function; it considers the semantic link between the terms by using ontology as external resource. We exploit in our case WordNet [7]. There are several methods to calculate semantic relations in WordNet. We chose to use the mathematical measure used in our previous model [19]. This work came to offer a solution for the process of semantic web services discovery according to user needs formulated under a request established in several terms. This allows calculating the length between two nodes representing two terms T_1, T_2 in WordNet ontology.

Modifying context is very important in our process; it decreases false answers and improves the overall quality of results. Our method ASWSCC is based on a context ontology [20] built and enriched with contextual data captured from different users. It allows us to consider changes in context in close moments during a single user session, dialing plans corresponding to the current use context are offered to user, respect to other plans related to

 ISBN: 978-1-61208-280-6

different instances are stored in a cache register during a change of state plans will be offered to him.

We use BPEL4WS (Business Process Execution Language for Web Services) specification or simply BPEL to express coordination between web services in the various plans proposed. It is the most industrially supported and well accepted by developers. In addition, there are several development tools that can help us in our work. For all these reasons, BPEL is the language that we have chosen for the implementation of this project. The result of the previous step is a set of execution plans in the form of finite automaton (state-transitions), states represent the different web services participating in the implementation plan, and transitions denote the semantic relationship between them.

For this reason a transformation "model to code" is necessary to generate the BPEL process. The work of Dautriche and Melliti [17] allows automatic generalization of BPEL code from an automaton and files corresponding WSDL service descriptions belonging to the automaton that describes the execution plan.

For this they have a tool for modeling behavior and then calculate the process model based on automaton. From this automaton a BPEL code generation is made. This work proposes a BPEL process executable generated by a tool developed in Java language.

To solve the problem of web services adaptation to change of user and service context, these services are invoked in a BPEL process, we have proposed an approach based on aspect weaving. We have presented a model for adaptable service composition characterized by adding two components, namely the context manager and aspect manager [5]. These are implemented as a web service, to allow the opportunity to easily integrate them in the same BPEL process where are services involved in adaptation operation.

We also proposed a model of context and we used a comparison algorithm of context variables based on a similarity computation algorithm proposed in the literature [18].

IV. CASE STUDY AND COMPARISON

To compare different service composition approaches, we realize an implementation of a process for web services composition; these services belong to different business domains (domain of travel organization which includes travel companies, hotels, restaurants, car rentals, and services organizing activities, library domain for purchase and rental services of books, DVDs and movies). Descriptions are published in a directory to provide best use and ease of access. We cite the example of travel organization agency which typically provides web services for consultation, reservation, payment and cancellation of travel tickets, hotel rooms, rental cars, restaurant reservations and organizing activities (Table I).

TABLE I. EXAMPLE OF WEB SERVICE DOMAIN AND ITS IMPUT AND OUTPOUT

WS domain	Input	Output
WS-Ticket booking	City departure, city arrival, date of departure and return date.	Code, Type, Price, city arrival, city arrival, departure and return date.
WS-Hotel	City, Input date, Output date.	Hotel Name, Address Hotel Class hotel, room type, duration and prices Hotel
WS-Restoration	Number of persons	Restaurant name, address, restaurant, restaurant class, many people, price
WS- Car rental	Duration rental, rental location	ID tenant, tenant Location, type and vehicle for hire
WS-Activities	City, Number of persons	List activities, duration, Region, activity price
WSCalculate order	List price	Total price
WS-Bank	Total price, payment, method, customer details	Customer details, Total price

Therefore, to provide these web services to its users, the travel agency must establish links with other companies: companies (airlines, agency of bus and train), car rental companies, hotel networks and catering, activities services that offers activities available in a given city. A bank is also required to facilitate financial transactions between users and agency organizing travels, or between the agency and other partners.

A. Experimentation

The prototype was developed under the operating system Windows Vista with open source graphics tools developed in Java and J2EE technology. We opted for Apache Tomcat application server that acts as JSP container, which allows its connection with a web server to deliver dynamic content to clients. We chose lexical database WordNet 2.0 as the global ontology, to compare terms and give similarity measure between concepts. We used the free dictionary Atla [8], which allows to translate words in French and whose target language is English, using its text files included with the dictionary.

The directory publication of our web services has been implemented in a relational database for a more explicit treatment of data; the conceptual model formalism is under « entity association». The service is described by the following functional parameters (category, function, input, and output).

We enrich this description with the context parameter named ContexteService will be represented with a separate table with the following view: (ContexteServiceCode, serviceID, Attribute, Value).

The various attributes ContexteService belong to the space defined in Figure 1. We have a database of 1,200 web services. We illustrate the functionality of our system by doing some individual images that show us steps of our Semantic Web services dynamic composition based on semantic matching and use context. We have two access rights (administrator and client). The administrator is the web service provider that connects to publish Web services (Figure 2). The client is the user who connects to use our process, he seizes his search query that can be focused on areas offered by our process mentioned above, the search

 ISBN: 978-1-61208-280-6

query is a complex task generic, and it will be treated to compose the existing web services to meet the needs of the user.

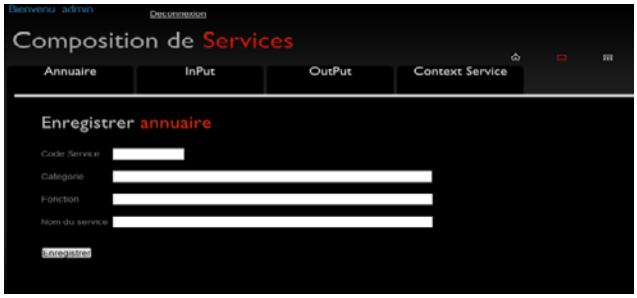


Figure 2. Administration Session

B. *Dynamic User Interface*

We opted for suggestion technique to help the user to edit his query; he starts typing his request, and a list box appears to propose him a list of tasks processed by our process loaded for a table « task».

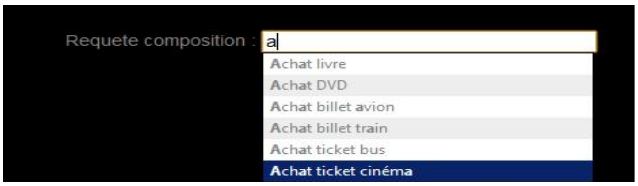


Figure 3. Menu suggestion

The user page also represents a source of contextual data collection, it gives the possibility to enter the type of device, its access mode, and through the IP address we can locate the network where the user is logged and can therefore face to location parameter.

This interface is 100% dynamic, information that appears in both fields of choice (device type, user type) are loaded from Context Utilization database, all instances of these two classes are extracted and offered to user.

This database is scalable and enriched by a domain expert to provide a broad definition of context. Use context is also about user preferences. According to his request, additional information must be completed.

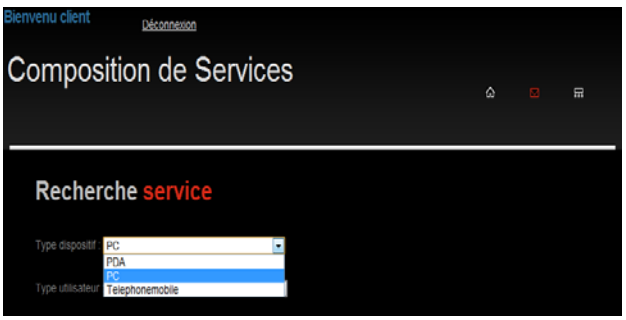


Figure 4. Collection of Context

Our user interface makes visible and invisible some input fields according to the request made. Example, for a query Q1 = « travel organization» four boxes "hotel, restaurant, car rental, activities" appear after query formulation. For a query Q2 = « buy book» a checkbox "service pack" appears to propose to user if he wants or not packing the purchased item. This is done to make our dynamic interface based on a database tasks interconnected through a semantic link.

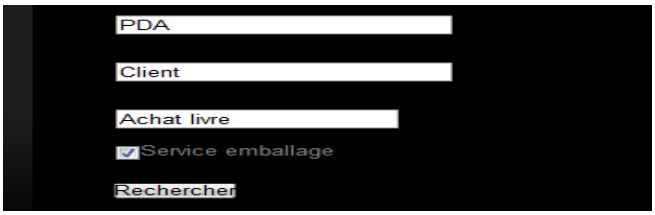


Figure 5. Dynamic Interface

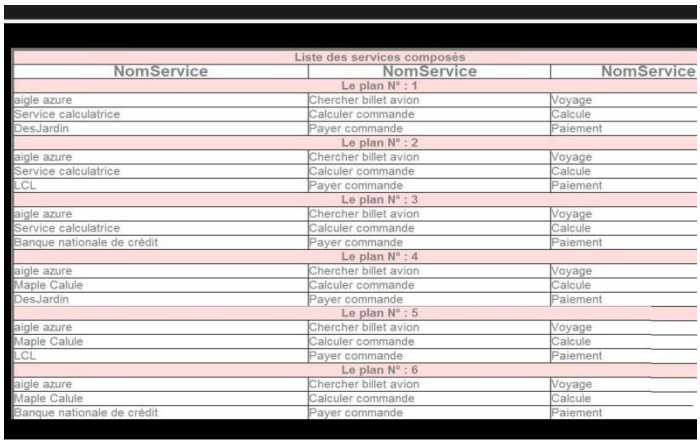
Our composition system begins to execute by decomposing user query in simple tasks using a knowledge base representing a composition pattern and discover web services by basing on their functional parameters, and use context, which reduces false answers. This process constructs and proposes then plans in order of composition sub queries in the pattern by matching input-output of services. This matching issue is based on domain ontology. At the end, we show to user the final result in a set of plans, each plan is a set of services. For context C1 = (PC, Client, IP @) the result of the query Q1 composition has 6 plans; see Figure 6.

Context change on one of C1's parameters implies a change in the proposed plans. For example, if the contextual situation changed C1 to C2 = (PDA, visitor, @ IP), the result is shown in Figure 7.

Some services are only available to customers and do not display the result on a type device «PDA». Through this example, we have seen that the steps proposed in our prototype composition gave us satisfactory results.

V. ASPECT WEAVING TOOLUSING THE TEMPLATE

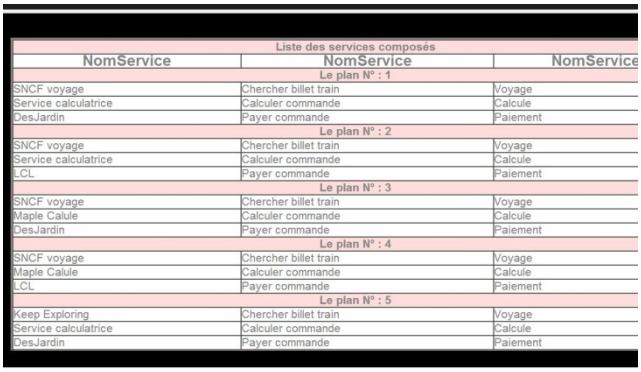
Changes can affect the context of web services during their executions and the need to adapt their dynamically becomes increasingly crucial.



Liste des services composés

NomService	NomService	NomService
Le plan N° : 1		
aigle azure	Chercher billet avion	Voyage
Service calculatrice	Calculer commande	Calcule
DesJardin	Payer commande	Paiement
Le plan N° : 2		
aigle azure	Chercher billet avion	Voyage
Service calculatrice	Calculer commande	Calcule
LCL	Payer commande	Paiement
Le plan N° : 3		
aigle azure	Chercher billet avion	Voyage
Service calculatrice	Calculer commande	Calcule
Banque nationale de crédit	Payer commande	Paiement
Le plan N° : 4		
aigle azure	Chercher billet avion	Voyage
Maple Calule	Calculer commande	Calcule
DesJardin	Payer commande	Paiement
Le plan N° : 5		
aigle azure	Chercher billet avion	Voyage
Maple Calule	Calculer commande	Calcule
LCL	Payer commande	Paiement
Le plan N° : 6		
aigle azure	Chercher billet avion	Voyage
Maple Calule	Calculer commande	Calcule
Banque nationale de crédit	Payer commande	Paiement

Figure 6. Result of composition before context change



Liste des services composés		
NomService	NomService	NomService
Le plan N° : 1		
SNCF voyage	Chercher billet train	Voyage
Service calculatrice	Calculer commande	Calcule
DesJardin	Payer commande	Paiement
Le plan N° : 2		
SNCF voyage	Chercher billet train	Voyage
Service calculatrice	Calculer commande	Calcule
LCL	Payer commande	Paiement
Le plan N° : 3		
SNCF voyage	Chercher billet train	Voyage
Maple Calule	Calculer commande	Calcule
DesJardin	Payer commande	Paiement
Le plan N° : 4		
SNCF voyage	Chercher billet train	Voyage
Maple Calule	Calculer commande	Calcule
LCL	Payer commande	Paiement
Le plan N° : 5		
Keep Exploring	Chercher billet train	Voyage
Service calculatrice	Calculer commande	Calcule
DesJardin	Payer commande	Paiement

Figure 7. Result of composition after context change

To solve the problem of the adaptation of web services invoked in a BPEL process to change of the client context and web service context, we have proposed an approach based on the aspect weaving. Its originality is the addition of two components namely the context manager and appearance manager [5]. To illustrate our model adaptation composition of semantic web services based on the weaving of aspects we are working on a concrete example of use. We have chosen the example of a process of buying books.

This process begins with the search and purchase books that result in a decision-command followed by the calculation of the amount of the order and finally the delivery of this order.

The BPEL processes represent complex service consists in this case of all three web services and the two components that will ensure adaptation where the context of the client or the web service change during process execution.

VI. COMPARISON WITH OTHER APPROACHES

We present now a comparative table between our approach and existing approaches explained above. This table is created according to some evaluation criteria to justify, evaluate, and positioning our proposal in comparison with other methods and techniques (see Table II).

TABLE II. COMAPARISON OF SERVICE COMPOSITION APPROACHES

	Manual Method	**Workflow based method**	**Planning based method**	**Our approach**
Automation level	Null : No automated process composition	Tool: the search and link to services are only automatic	High : Practically, all the process is done automatically	High : Practically, all the process is done automatically
Dynamic layer	All is done statically	Process model is done statically. Link to concrete services is dynamic	The generation of composition plan is done dynamically	High : Interface user 100% dynamic. The generation of a plans of composition is dynamic
Formulation of the query	Very low level : the user that defines the composition plan	Low level : the user should implement the process model using flow language	Specifies the initial state, the final state, and with certain constraints mixture of very low level (logic language)	The scheme of composition is defined as sub queries using a knowledge base, each sub query models of semantic web services
Layer of Granularity of generated composition plans	One generated plan (executed plan)	Abstract Workflow and executable workflow	One generated (executed plan)	Many composition plans that are generated
Used supports and abstraction level	No semantic and contextual support l (UDDI registries)	Using of syntactic descriptions	Using of logic formalisms	Using ontologies. Extraction and management context. Enriching service descriptions with context
Discovery web service	Primitive (in the case of UDDI registries)	Using a research motor UDDI	In the profiles DAML-S or OWL-S	Dynamic discovery based on semantic matching, taking into account the context of use.

We identify six evaluation criteria to be considered (first column). Our approach uses AOP to compose services dynamically. This approach depends also on mathematical formula of semantic matching algorithm [19].

Compared to other models listed in the table below, and with respect with these criteria, our approach has demonstrated automation in all the process steps, its trigger element is the user query.; It is inputted in a dynamic interface changing according to the user's needs. This request can be divided into several simple tasks based on a decomposition model proposed and explained in our anterior work [5]. This process has been also proof of wealth due to a semantic layer that contains all the necessary data to complete the semantic web services discovery. These web services will belong to the composition plan. This discovery process is based too on a parameter that we consider very important: the context. In the end several composition plan are available to the user.

VII. CONCLUSION AND FUTURE WORKS

Web services adoption is a major advance in the development of interoperable information systems. In particular, the composition of services can meet the needs

increasingly complex of user, by a combination of web services within a single business process. However, despite this widespread adoption of Web services, many obstacles prevent their reconciliation in the composition, or may occur within a BPEL process in a state change, the context for example. To solve this problem of the Web service adaptation invoked in a BPEL process and to satisfy the client context and web service context, we have proposed approach based on the aspect weaving, its originality is the addition of two components namely the context manager and appearance manager. These two managers are discussed in detail in Mcheick et al., 2012 [5].

In this paper, we presented the detailed steps (method) we followed to model semantic Web service composition. Then, this composition method is compared briefly with other service composition methods such as manual, workflow and planning methods. This comparison shows that our approach has many advantages in terms of automation and dynamic level, layer of granularity of generated composition plans and others.

This platform needs more investigation in terms of semantic layer and comparison with all the platforms of service composition, such as METEOR-2 [11], SELF-SERV [9], and SHOP2 [13]. As of our perspectives, we need to consider more applications to measure the satisfaction of user requests based on the context and weaving aspects.

ACKNOWLEDGMENT

This work was sponsored by the University of Quebec at Chicoutimi (Quebec), Canada.

REFERENCES

[1] B. Benatallah, R. Dijkman, M. Dumas and Z. Maamar, “Service Composition?: Concepts, Techniques, Tools and Trends,” In: Z. Stojanovi and A. Dahanayake, Eds., Service-Oriented Software System Engineering: Challenges and Practices, Idea Group, pp. 48-66, 2005.

[2] M. Keidl and A. Kemper, Towards context-aware adaptable web services, Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters, May 19-21, New York, NY, USA, doi:10.1145/science.1013367.1013378, 2004.

[3] T. Chaari, F.Laforest, and A. Celentano, “Design of Context-Aware Applications Based on Web Services,” LIRIS. Dipartimento di Informatica. INSA Lyon, France, 2005.

[4] M. Vukovic and P. Robinson, “Adaptive, planning based, web service composition for context awareness,” Proc. 2nd International Conference on Pervasive Computing, Vienna, Austria, June 2004.

[5] H. Mcheick and A. Hannech, “Modèle de composition des services web sémantiques par orchestration à la demande,” Proc. of IEEE Canadian Conference on Electrical and Computer Engineering (CCECE 12), April 29-May 2, Montreal, Canada, 2012.

[6] A. Dey, G. Abowd, and D. Salber, “A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications,” Human-Computer Interaction Journal, vol. 16, issues 2-4, pp. 97–166, 2001.

[7] http://fr.wikipedia.org/wiki/WordNet, [retrieved : 04, 2013].

[8] http ://atla.revdanica.com/, [retrieved: 05, 2006].

[9] Q. Sheng, B. Benatallah, M. Dumas, and E. Mak, “SELF-SERV: A Platform for Rapid Composition of Web Services in a Peer-to-Peer Environment,” Proc. of the 28th International Conference on Very Large Data Bases (VLDB 2002), Aug., Hong Kong, China. Morgan Kaufmann 2002, pp.1051-1054, 2002.

[10] B. Benatallah, Q. Sheng,, and M. Dumas, “The SELF-SERV environment for Web services composition,” IEEE Internet Computing, vol.7, n°1, pp.40-48, 2003.

[11] N. Oldham, C. Thomas, A. Sheth, and K. Verna, “METEOR-S Web Service Annotation Framework with Machine Learning Classification,” Proc. of the 1st International Workshop on Semantic Web Services and Web Process Composition, SWSWPC 2004 (SWSWPC 2004), San Diego, CA, USA. LNCS Springer, 2004, pp.137-146, 2004.

[12] R. Aggarwal, K.Verma, J. Miller, and W. Milnor, “Constraint Driven Web Service Composition in METEOR-S,” Proc. of the IEEE International Conference on Services Computing (SCC’04), Sept. 2004, Shanghai, China. IEEE Computer Society, pp.23-30, 2004.

[13] E. Sirin, B. Parsia, D. Wu, J. Hendler, and D. Nau, “HTN Planning for Web Service Composition Using SHOP2,” Journal of Web Semantics, vol.1, n°4, pp.377-396, 2004.

[14] L. Cabral,, J. Domingue, S. Galizia, A. Gugliotta, V. Tanasescu, C. Pedrinaci, and B. Norton, “IRS-III: A Broker for Semantic Web Services Based Applications,” Proc. of the 5th International Semantic Web Conference (ISWC 2006), Nov. 2006, Athens, GA, USA. LNCS Springer, pp. 201-214, 2006.

[15] Y. Charif and N. Sabouret, “Coordination in Introspective MultiAgent Systems,” Proc. of the International Conference on Intelligent Agent Technology (IAT’07), pp. 412-415, Silicon Valley, California, USA, 2007.

[16] F. Pourraz, “Diapason une approche formelle et centrée architecture pour la composition évolutive de services Web,” Thèse de Doctorat, LISTIC : Laboratoire d’Informatique, Systèmes, Traitement de l’Information et de la Connaissance, Dec. 2007.

[17] A. Dautriche and T. Melliti. “Génération automatique de code pour services web composites,” Rapport de TER, 2010.

[18] S. Lajmi, C. Ghedira, K. Ghedira, and D. Benslimane, “WeSCo CBR: How to compose Web Services via Case Based Reasoning,” e-Business Engineering ICEBE, pp. 618-622, 2006.

[19] H. Mcheick, M. Adda and A. Hannech. “Web Service Discovery Model Based on Context,” American Journal of Software Engineering and Applications issued by Science Publishing Group, Vol.1, No.1. December 2012.

[20] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. Mcllraith, S. Narayanan, M. Paolucci, B. Parisia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara, “Owl-s : Semantic markup for web services,” Technical report, W3C, 2004.

[21] F. Curbera, Y. Goland, J. Klein, F. Leymann, and S. Weerawarana, “Business Process Execution Language for Web Services (BPEL4WS),” Version 1.1, May 2003.

[22] http://www.ibm.com/developerworks/webservices/library/ws-bpelcol4/, [retrieved: 11, 2002].

[23] T. Chaari, F. Laforest and A. Flory “Adaptation des applications au contexte en utilisant les services WEB ” UbiMob 2005: 111-118

 ISBN: 978-1-61208-280-6

Towards a New Trust Model for Health Social Networks

Sojen Pradhan and Valerie Gay
Faculty of Engineering and IT
University of Technology Sydney (UTS)
Sydney, Australia
e-mail: {Sojendra.Pradhan,Valerie.Gay}@uts.edu.au

Abstract—**More specific purpose driven social networking sites have emerged since social networking sites have gained popularity by bringing people with shared interests together to interact. In health care, they are referred as Health Social Networks (HSNs). Given the sensitive nature of health information, trust is fundamental for them. The emergence of pervasive and ubiquitous computing environment and overwhelming information available online is helping the health social networking sites gain popularity at a fast pace. Health social networkers are willing to create, share or retrieve trustworthy health or lifestyle related information. Therefore, it is essential that trust is stipulated and scrutinised to understand how the users perceive healthcare, how they decide to interact with HSNs. This paper analyses external factors such as perceived information quality, perceived system quality, perceived reputation and perceived trust signs which impact the trust model for HSNs. In particular, 'perceived reputation' based on the factor 'who has recommended the site' is given an emphasis on this paper. It highlights that popularity of social networking sites is changing the way trust models have been defined in the past. This is because social relationships created via social networking sites are also impacting on choosing the HSNs and how users are sharing health information on these platforms.**

Keywords-Online health information; social networking sites; health social networks; trust model.

I. INTRODUCTION

The use of SNSs (social networking sites) has made a substantial impact on the revolution of health care digital communication. Health information is generally sourced from health care professionals but an increasing number of healthcare consumers turn to websites and SNSs nowadays for the source or second opinion. Due to both pervasive and ubiquitous nature of ICTs (information and communication technologies), the number of people sharing health information online and the number of social networking sites for health-related information is increasing [1,2,3]. One of the popular topics for people to participate and share online is health-related information. Health information is shared with other users although personal health information is considered to be sensitive. It is one of the basic characteristic of human being that when we experience something positive, we tend to share with peers and recommend others. For example, if we watch a good movie, we tend to recommend this movie to our friends to watch. In a social networking environment, when a user shares information on a particular topic, they discuss their positive experience and subliminally recommend, validate and endorse the experience. In this paper, we use the term 'HIs' (health infomediaries) for providers who provide unbiased health information online to users, through which they have choice to make their health related decisions [4]. Another term 'HSNs' (health social networks) is used to cover social networking sites where users search, self-track, discuss health and lifestyle or fitness related information [2, 5].

While there is an increasing trend in using the ICTs to search and communicate health information online, the demand for high quality information has also been rising [6]. At the same time, there is also a legitimate concern for security and privacy. The impact of social networks on healthcare is the subject of studies [5] and there are serious concerns as the healthcare consumers rely on the information provided by the health related platforms such as HIs, HSNs, Apps. Despite the warnings 'not to use the information without consulting a health care professional', consumers use the information to make health related decisions. Therefore, questions such as, 'how do health consumers know the platform is trustable and the provided information is well researched? Are the health social network sites safe?' have risen. Such platforms may have been created to provide information to facilitate sales of a product or service [7] or capture private information in exchange of perceived benefit [3]. Trust plays an important role for healthcare consumers to reduce uncertainty in technology-mediated environment [8].

There have been many studies regarding trust on the websites but only a few researchers have focused on the health related online information. Among them, Song and Zahedi [4] suggested that the quality of information and the level of trust the healthcare consumers have with the health platforms are very important to make their health decisions. It has been argued that the trust would be more reliant on the content for health related information, than other factors such as how it is presented, HCI (Human Computer Interaction) factors or the credibility of the platform. Less priority seems to be given to the factor such as 'who has recommended the site for the particular health information or to exchange health related information'. Pew Research Center [9] has reported that 80% of Internet users in US have looked for information about health topics or similar health issues they are facing. It has also been reported that over 3,000 hospitals have social networking sites which includes over 700 Facebook pages [10]. However, due to the sensitive nature of the health related information, the health social network

 ISBN: 978-1-61208-280-6

users may disclose information and describe terms which could be misleading or misinterpreted. This certainly creates some challenges to the health social networkers.

This paper is a step towards a new trust model for HSNs. Section 2 defines HSNs and the risks associated with them. Section 3 focuses on trust and discusses its vital role in reducing those risks. Section 4 analyses external factors affecting a concept of trust. Section 5 proposes a new trust model for health care recommendation systems. The paper concludes with open issues and future work.

II. HEALTH SOCIAL NETWORKS AND RISKS

Since SNSs have been gaining popularity, more specific purpose driven social networks have emerged in addition to the popular sites such as Facebook for general purposes and LinkedIn for career specific. Healthcare consumers (both professionals and consumers) have moved from searching information online to sharing information and in fact interacting with other users within the platforms [11]. They are able to find other users in similar illnesses or health situations and interact with each other about their conditions, symptoms and treatments in the sites like PatientsLikeMe, DailyStrength and many others [1]. This environment provides great opportunities for healthcare consumers to be able to connect and relate with each other [11]. It has been reported that 23% of chronic health e-patients with cancer, diabetes, or heart disease have searched for other patients with similar conditions [11]. Other studies such as 'Point of Care' Survey conducted by Wolters Kluwer Health revealed that the physicians have changed initial diagnosis of patients based on new information accessed online resources [12].

With the increased number of healthcare consumers turning to HSNs for retrieving and sharing health information, the number of the users who rely on the information from these platforms is rising. It raises the potential danger of using the health information incorrectly by healthcare consumers in a short or long term. The degree of danger unequivocally depends on the skills and knowledge of the healthcare consumers, such as understanding of medical or scientific vocabulary and biomedical knowledge, to interact with the HSN communities and other health related platforms [2]. BetterHealthChannel [13] has listed some of the potential risks associated with health information online. These include, wrong diagnosis, misunderstanding of medical jargons, self-medication may delay visit to the health professional and hence miss out on appropriate early and appropriate treatment for the illness, a delay may cause serious complications or death, and may have unwanted side effects or interact with other medications.

III. 'TRUST' AS A MAJOR FACTOR

Trust is a very complex phenomenon. There are many definitions and studies of trust in many aspects of lives. Many experiments and surveys have been conducted and developed trust models accordingly. Yet there is no universal definition of trust that everybody can share and the concept of trust remains elusive [14]. A simple reason for that is that 'trust' has numerous and diverse meanings. On a daily context, 'trust' is a term with many meanings [15]. "Trust is an important lubricant of a social system", because it can enhance efficiency [16]. It has been shown how trust as a high level of altruism can increase efficiency of people working together [17]. In general terms, trust is a relationship between the trustor and the trustee. In the context of health information; the trustor is a health social networker (healthcare consumer) and the trustee is the HSN platform.

Many researchers and scientists have defined and categorised many different types of trust. Among them, Josang et al. [18] used Reliability trust and Decision trust in the context of health information. The measurement of reliability trust is to provide the best health-related information based on ability, knowledge, skills and competence of the trustee (platform or information provider). This could be determined by the credibility, qualification and history of successful stories or case studies provided. Decision trust can be measured based on the actual actions the users take after getting exposed to the information in the platform. This could be influenced by the circumstances the users are in, for example the urgency of the need, or most importantly who has recommended this platform.

Many researchers emphasised the importance of initial trust for users that attract them to visit a platform for the first time. Song and Zahedi [4] designed a trust model and not only focused on the initial trust component for health related platforms, but the dynamics of trust revision as per time of loyal users. Time is important component on measuring trust because the level of trust may increase or decrease over time.

Adams [19] focused on reliability issues in the context of interaction of health consumers with information in the technology-mediated environment. More specifically on the quality of information (credibility and accuracy) and the healthcare consumer's behavior in terms of creating, exchanging and retrieving within social network environment. Quality of information is not just about ratings of the health information available online [20], but the credibility of the content in line with the concept of reputation and a collective measure of trustworthiness [19]. Reputation building is prevalent within SNSs and recommendation sites as these platforms provide the opportunities to reach more consumers and facilitate to create, share and retrieve information online. In hindsight, the reputation building process can be manipulated through pre-formatted templates, which could lead to suggest specific products or services [19].

With the increased number of healthcare consumers interacting on HSN platforms, more issues about reliability and trustworthiness will be encountered in making decisions for their health issues. In this paper, an existing trust model for health infomediaries by Song and Zahedi [4] is reviewed. In addition to existing external factors, this model will be altered to emphasize impacts of social influence such as 'who has referred to this particular HSN platform?' Later on, a new trust model is proposed.

 ISBN: 978-1-61208-280-6

IV. NEW TRUST MODEL FOR HSNS

The model is based on the framework of TRA (Theory of Reasoned Action) and conceptual trust model designed by Song and Zahedi [4]. TRA has five main components: 'external factors', 'trust beliefs', 'trust attitudes', 'intentions', and 'behavioural outcome' [21]. These five components lead to develop relationship between the trustor (health social networker) and trustee (HSN platform) and the relationship could be either positive or negative. The five components and the 'relationship development' steps are shown in the Fig. 1.

While making decisions, external factors assist to outline the trust beliefs formation. This is what influences the formation of trust attitudes and then intentions towards determining eventual behaviour such as whether to act or not on the information provided or extracted from the HSN platforms for their health issues. TRA considers that trust beliefs lead to trust attitudes, and then they lead to behavioural intentions and becomes behaviour [22].

The first three components from the TRA framework: 'external factors', 'trust beliefs' and 'trust attitudes' are further refined into a conceptual framework and shown in Fig. 2.

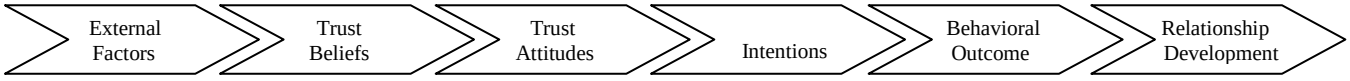


Figure1. Components of e-commerce exchange relationship development framework [21]

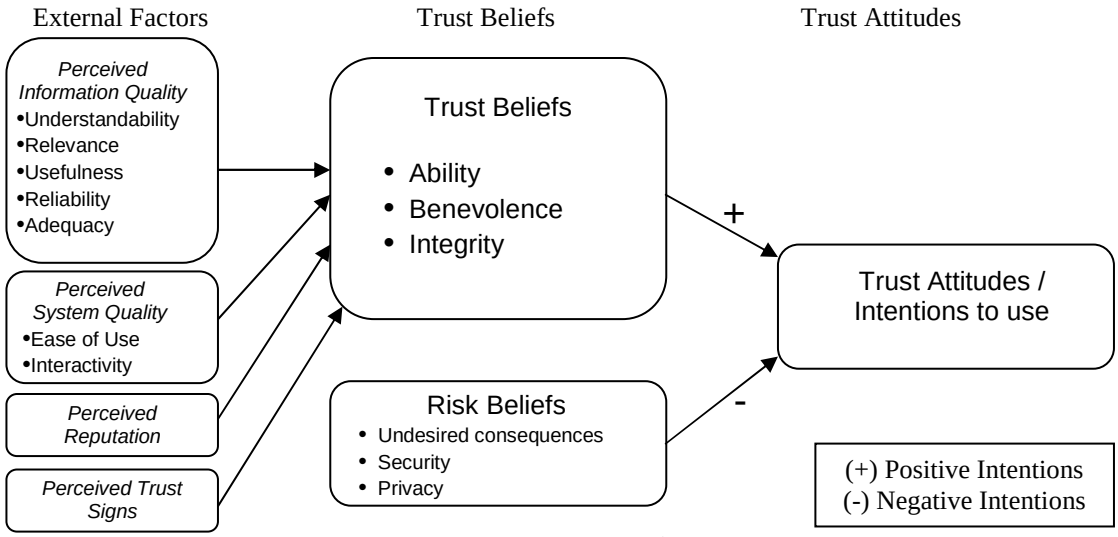


Figure 2. Conceptual model of trust [4]

The intentions to use health information or take part in interaction or exchange of personal health information in HSN platforms are derived through the level of trust to the particular platform. If the intentions are positive and the experience turns out to be a favourable one, this could lead to the development of a relationship with the HSN platform (trustee). Inherently, health consumer would likely re-visit the platform and recommend the platform to others. This is where power of SNSs comes into play. In terms of gaining the trust, the first impression or initial trust is very important. Through a good initial trust, the users will be willing to use and share health information within the HSN platform. Therefore, social influence 'who refer to the site' has a critical role. Once the trust is established, there will be more interactions between the trustor and the trustee over time. The level of trust is determined by the information quality (credibility), system quality and satisfaction to the trustor which will be developed over the time [23].

A. *Perceived Information Quality*

The measurement of information quality is evolving with the pervasive ICTs in the healthcare domain. In the model we selected, perceived information quality has been further classified into the following sub-categories:

- Understandability: Understandability means clarity of the information to the user. Medical and scientific vocabulary could create challenges or users to understand. As long as the HSN platform is destined for the general public, the trustor needs to be careful of the vocabulary used. There is always a danger that users may misunderstand the terminologies used.
- Relevance: Relevance refers to the appropriateness of the information to the users. If the information is understood, the users are able to verify whether it is relevant to their needs. Medical or health related knowledge is important to understand the relevance.
- Usefulness: Due to the sensitive nature of health information, healthcare consumers are concerned about

any form of digital communication. Perceived usefulness (PU) of the extracted information influences the trust beliefs to a positive territory and eventually influences behavioural intentions [24].

- Reliability: Reliability refers to the credibility of the trustee and the accuracy of the information. It is a broad terminology and may well incorporate technical aspects of the platform and health consumer behavior [19].
- Adequacy: Adequacy refers to the completeness and references provided. Completeness means an extensive coverage of health-related information on the specific topic. This could portray as a sign of the commitment of the platform to the users by providing unbiased information and references.

Besides these sub-categories, there are many other important factors to qualify health-related information quality such as timeliness, accuracy, clarity and so on. They are not covered in the model in the Fig 2.

B. Perceived System Quality

A study was carried out to determine overall satisfaction with system quality and information quality for health information. It was reported that system quality (usability) played a greater role than information quality in the study [25]. Both perceived ease of use (PEOU) and perceived usefulness (PU) are basic ingredients to support the technology acceptance model (TAM), an information system theory which models how users accept and use a technology. This theory later forms the trust antecedents of intentions to utilise the health-related information, as intended at the time of creation [26].

- Ease of use: The ease of use refers to the usability of the HSN platform which will determine whether the users want to spend time on it. Perceived ease of use (PEOU) influences the perceived usefulness (PU).
- Interactivity: Interactivity refers to the web features that ease the user's experience for the search and potentially even personalize the information based on the search criteria.

C. Perceived Reputation

The term reputation can be defined as the social influence of trust, which can be referred as social exchange theory that define one party's reputation based on a third party's ability to tell stories about its trustworthiness. The terms 'reputation' and 'trust' are strongly linked to each other. Reputation is usually influenced by the past behaviour. A repeated visit and prior positive experience of health consumer with the platform denotes the perceived reputation. Any good or bad experiences or result is easily circulated via social networking sites instantly. It is even more important for health-related information to be distributed faster if more people to get benefit or protect from.

D. Perceived Trust Signs

Trust signs are used to reassure the healthcare consumers that there are no risks associated with HSN platforms to interact to or retrieve information from and reinforce integrity of the provider. The use of trust signs is necessary to convince the users that they can trust the HSN platform and its information [4].

Deshpande and Jadad [20] provided five broad categories to evaluate the quality of online health information and depicted as trust signs:

- Codes of conduct (e.g., Australian Medical Association),
- Quality labels (e.g., Health on the Net Foundation [HONcode]),
- User guides (e.g., DISCERN Online),
- Filters (e.g., intute.ac.uk) and
- Third party certification (e.g., Hi-ethics, Utilization Review Accreditation Commission [URAC])

Overall, the external factors from the conceptual model in Figure 2 influence the trust beliefs and ultimately influence the HSN platform user's intention to act on the information extracted from HSN platforms. If the perceived information quality, perceived system quality, perceived reputation, perceived trust signs and satisfaction are all positive, the HSN platform users will come back to retrieve and share health information in the HSN platform and in fact recommend to others [27]. Loyalty is critical to sustain the systems however these information tend to have temporal effect as soon as the user receive the required information, there is no incentive for them to come back. The user satisfaction is what makes the users loyal and recommend the system to others.

Intention of using the information extracted from the HSN platform is significantly relied on the urgency of the matter, need and circumstances of the person at the time and the trust is crucial for these circumstances. From the conceptual trust model, the impact of the external factors, specifically the influence from the third party based on who has recommended visiting the HSN platform, will be researched further. This factor is very dynamic and very complicated to measure as these variables would change from case to case.

V. PROPOSED MODEL

The conceptual trust model analysed external factors which affects trust in the health information and the health related platforms. However, in this research, we focus on the perceived reputations, one of the external factors and specifically, impact of 'who recommended the platform or information?'

Since the explosion of social media, more information is being shared online. The social behavior of human being has been replicated in social reviews sites or recommendation sites by allowing more users to interact and share their experiences in an unbiased environment. Depending on who recommended the health platform or information provider, users have tendency to follow through better. If the experience is good and satisfied during the process, they will tend to continue to use and recommend to others further. Based on this, a new trust-based model for dynamic

 ISBN: 978-1-61208-280-6

healthcare recommendation system is proposed as shown in the Fig. 3.

This model has 3 steps. Firstly, users or patients will have some preferences (criteria) while searching for health-related information or healthcare providers. The criteria such as location, symptoms, age, specialist, availability and others are used. Based on these criteria, a system would provide a list of health-related information or providers. Which one in the list to use in this information overloaded age?

Because people tend to rely more on recommendations from people they trust, we would evaluate trust within their own social networks which can help to sort out the list from the previous step. The trust is what would influence how and what information is going to be used by the users. There are many ways to evaluate trust within SNSs. The evaluation of trust in our research will be done by analysing the strength of relationships among users in the social network. The influence is directly proportionate to strength of relationships. Analysis of similarity in the context is another element we will focus on, such as symptoms, side-effects, and behaviours among the users. In addition, influence could also be determined by how knowledgeable the trustee in the specific healthcare area is. If the person is an expert in the area, his/her opinion will be given more priority by the trustor. Analysing these information, a trust value would be generated. Based on the trust value, the health information or the healthcare provider would be selected.

In some cases, the users or patients are able to verify the information (or provider) further with the existing online information (crowdsources) to assure that the information is trustworthy. It is the last step of the proposed model, which is an alternative, because the information may not be available for all information (or providers).

This model accommodates the users' preferences (criteria) and users' trust within their own network to be able to filter through to the best possible result while looking for health-related information or healthcare providers.

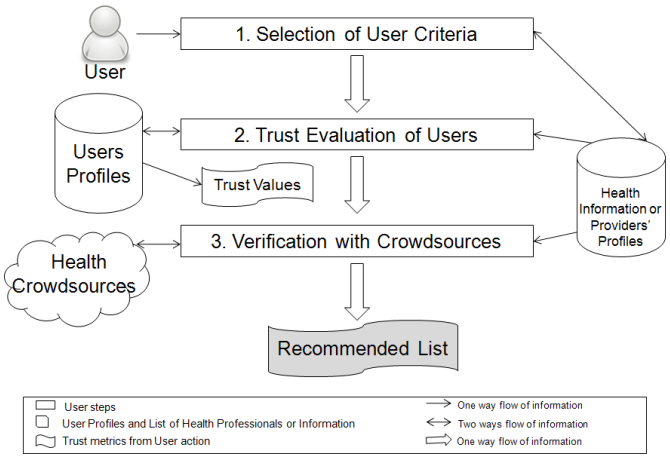


Figure 3. Proposed Model

VI. Open Issues and Future Work

SNSs are open platform for communication and they provide a meeting place to create and share experiences in life. Users share information with each other in the SNSs, regardless of whether knowing or not knowing the remote user, which provides both opportunities and challenges for sensitive health information. How do we know the information publicised on the HSNs or shared in the SNSs are accurate? Not only content but also the source of the information is very important. The search engines cannot provide whether the source is trustworthy or not. Trust has been regarded as one of the major factors the users consider in the process of searching and taking actions on health-related information. Yet, it is very subjective to determine the trust value as it is extremely dynamic and changes quickly with many dependent variables such as time, situation, knowledge, experiences and many others.

In the future work, we will focus on the impacts of health social networks to the trust model and test some hypotheses to prove the significance of the impacts on the model in specific health care areas such as dental care. We will divide trust into internal (local) and external (global) trust factors. Internal trust is generated within a user either through existing relationships in their network or through their sharing experience in a particular HSN. External trust will be through existing ratings and review sites about the health information online. Impacts of reviews and recommendations in conjunction with own level of trust to the health platform and a particular information provider as a trustee will be studied further in the domain.

References

[1] M. Swan, "Emerging patient-driven health care models: an examination of health social networks, consumer personalised medicine and quantified self-tracking", International journal of environmental research and public health, vol. 6, no. 2, Feb. 2009, pp. 492-525, doi:10.3390/ijerph6020492.

[2] Y.W. Webster, E.R. Dow, J. Koehler, R.C. Gudivada, and M.J. Palakal, "Leveraging health social networking communities in translational research", Journal of biomedical informatics, Feb. 2011, vol. 44, pp. 536-544, doi: 10.1016/j.jbi.2011.01.010.

[3] J. Williams, "Social networking applications in health care: threats to the privacy and security of health information", 2nd Intl Workshop on Software Engineering in Health Care at ICSE, May. 2010, pp. 39-49.

[4] J. Song and F.M. Zahedi, "Trust in health infomediaries", Decision Support Systems, vol. 43, no. 2, Jan. 2007, pp. 390-407, doi: 10.1016/j.dss.2006.11.011.

[5] V. Gay, and P. Leijdekkers, "The Good, the Bad and the Ugly About Social Netwoks for Health Apps", 9th IEEE/IFIP Embedded and Ubiquitous Computing (EUC), Oct. 2011, pp.463-468, doi: 10.1109/EUC.2011.69.

[6] N. Pletneva, S. Cruchet, M.A. Simonet, M. Kajiwara, and C. Boyer, "Results of the 10 th HON survey on health and medical Internet use", Health on the Net Foundation, 2011.

[7] C. Boyer, V. Baujard, and A. Geissbuhler, "Evolution of Health Web certification, through the HONcode experience",

Studies in health technology and informatics, vol. 169, 2011, p. 53.

[8] B. Hernández-Ortega, "The role of post-use trust in the acceptance of a technology: Drivers and consequences", Technovation, vol. 31, Jul. 2011, pp. 523-238, doi: 10.1016/j.technovation.2011.07.001.

[9] S. Fox, "The social life of health information, 2011", Pew Research Center's Internet and American Life Project, [retrieved: April 2013, http://www.pewinternet.org].

[10] N.P. Terry, "Fear of Facebook: Private Ordering of Social Media Risks Incurred by Healthcare Providers", Working Paper, Aug. 2011,<http://works.bepress.com/nicolas_terry/2>.

[11] W.L. Lober and J.L. Flowers, "Consumer Empowerment in Health Care Amid the Internet and Social Media", Seminars in Oncology Nursing, Vol. 27, No.3. Aug. 2011, pp: 169-182, doi: 10.1016/j.soncn.2011.04.002.

[12] M. Merrill, "Docs turn to Google, Yahoo for health info, survey finds" Healthcare IT News, Nov. 2011, [retrieved: April 2013, http://www.healthcareitnews.com/news/survey-docs-turn-google-yahoo-health-info].

[13] Victoria-Government, "Health information on the Internet,", Better Health Channel, Jun. 2012, [retrieved: Apr 2013, http://www.betterhealth.vic.gov.au/bhcv2/bhcarticles.nsf/pages/Health_information_on_the_internet?open].

[14] N. Watanabe, "Role of trust in cross-cultural adaptation: the perspective of international degree students at a Finnish university", Master's Thesis, University of Jyväskylä, 2008.

[15] J.M. Sinclair and R. Carter, Trusts the text: language, corpus and discourse, Routledge, New York. 2004.

[16] F. Fukuyama, Trust: Human Nature and the Reconstitution of Social Order, Free Press, New York, 1996, p. 151.

[17] R. Chami and C. Fullenkamp, "Trust and efficiency", Journal of Banking & Finance, vol. 26, no. 9, Sep. 2002, pp. 1785-1809, doi: 10.1016/S0378-4266(02)00191-7.

[18] A. Josang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision", Decision Support Systems, vol. 43, 2007, pp. 618-644.

[19] S.A. Adams, "Revisiting the online health information reliability debate in the wake of "web 2.0": An inter-disciplinary literature and website review", International Journal of Medical Informatics, vol. 79, no. 6, Jan. 2010, pp. 391-400, doi: 10.1016/j.ijmedinf.2010.01.006.

[20] A. Deshpande and A.R. Jadad, "Trying to measure the quality of health information on the internet: is it time to move on", Journal of Rheumatology, vol. 36, no. 1, Jan. 2009, pp. 1-3, doi: 10.3899/jrheum.081101.

[21] A. Salam, L. Iyer, P. Palvia, and R. Singh, "Trust in e-commerce", Communications of the ACM, vol. 48, no. 2, Feb. 2005 pp. 72-77.

[22] D.H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology", Information Systems Research, vol. 13, no. 3, Sep. 2002, pp. 334-359.

[23] F.M. Zahedi and J. Song, "Dynamics of trust revision: Using health infomediaries", Journal of Management Information Systems, vol. 24, no. 4, 2008, pp. 225-248, doi: 10.2753/MIS0742-1222240409.

[24] R. Klein, "Internet-Based Patient-Physician Electronic Communication Applications: Patient Acceptance and Trust", e-Service Journal, vol. 5, no. 2, 2007, pp. 27-51.

[25] M. Bliemel amd K. Hassanein, "Consumer satisfaction with online health information retrieval: a model and empirical study", e-Service Journal, vol. 5, no. 2, 2007, pp. 53-83.

[26] V. Cho, "A study of the roles of trusts and risks in information-oriented online legal services using an integrated model", Information & Management, vol. 43, no. 4, Jun. 2006, pp. 502-520, doi: 10.1016/j.im.2005.12.002.

[27] J. Gummerus, V. Liljander, M. Pura, and A. van Riel, "Customer loyalty to content-based web sites: the case of an online health-care service", Journal of Services Marketing, vol. 18, no. 3, 2004, pp. 175-186, doi: 10.1108/08876040410536486.

 ISBN: 978-1-61208-280-6

A Web Service Migration Framework

M. Mohanned Kazzaz
Department of Information Systems
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
Email: ikazzaz@fit.vutbr.cz

Marek Rychlý
Department of Information Systems
Faculty of Information Technology
Brno University of Technology
Brno, Czech Republic
Email: rychly@fit.vutbr.cz

***Abstract*—Service-oriented software systems that operate in highly dynamic or mobile environments have to be able to adapt at runtime to changing environmental conditions. This includes not only adaptability of their individual services, but also ability of the whole systems as service compositions to preserve their integrity and to keep their best performance. This paper presents a concept of service migration in service-oriented architecture as an approach to enable the adaptation of service-oriented systems in changing environments. Moreover, a description of migratable services and service providers by means of migration conditions is proposed and used by service-oriented systems to keep functionality and quality of their services. Finally, the paper describes a prototype design of a framework for the service migration according to the migration conditions and to an user-defined migration decision strategy.**

***Keywords*—Web services; Quality of service; Adaptive systems.**

I. Introduction

In order to operate effectively in mobile environments, a software application has to be able to adapt at runtime to changing environmental conditions such as high volatility and fluctuation of available resources, variable quality of utilised services, unstable network topology, etc. In the case of a distributed component-based system, the changing environmental conditions means not only adaptability of the system's individual components, but also ability of the whole system to preserve its integrity and to keep the best performance [1].

Currently, information systems designed as the component-based systems often utilise service-oriented architecture (SOA) and Web service technology. The service orientation allows to decompose a complex software system into a collection of cooperating and autonomous components known as services. These services cooperate with each other to provide a particular functionality of the implemented system with defined quality.

This paper deals with service migration in SOA as an approach to enable the adaptation of component-based and service-oriented systems in highly dynamic and heterogeneous environments. While traditional models of SOA assume that services are provided permanently by service provides which are predefined at a system's deploy-time or found in service registries at its runtime [2], the service migration enables services to be transparently moved across various network nodes that act temporarily as service providers according to their availability and their resources. Through the service migration, the system is able to cope with the inherent environmental dynamics and to keep functionality and quality of its services (e.g., to employ temporarily available mobile devices as service providers, to react to possible failures of service providers with unreliable resources, etc.).

The rest of this paper is organised as follows. Section II describes a process of service migration including our logical model for migration decision process. The migration process is implemented in Section III as a framework for Web service migration. In Section IV, we review the main approaches that are relevant to our subject. Section V discusses advantages and disadvantages of the proposed framework, especially in comparison with other state-of-the-art approaches. Section VI outlines ongoing implementation of the framework and future research work. Finally, Section VII concludes the paper with summary of our work.

II. Process of Web Service Migration

The migration of a particular service should be considered if its provider is not able to guarantee the functionality or quality of the service and there is no alternative service or service composition that will match a semantic and qualitative description of the original service, i.e., that can provide the same functionality and required quality.

A. Migration Conditions and Migration Decision

We define two groups of *migration conditions*, i.e., the conditions which indicate the need of service migration. The first group contains predefined conditions concerning a provider's runtime state, e.g., network traffic, memory usage, CPU usage, and battery state. The second group consists of user-defined migration conditions, which can be used together with an user-defined migration decision strategy to fully customise a process of decision making for potential service migrations.

The *migration decision* whether start the migration of services provided by particular service providers and how to find target service providers (that will provide the services after the migration) is based on required quality of the services and on optimal and actual states of the service providers.

At the beginning, a migration controller is periodically gathering information about current state of each service provider (e.g., current battery state or resource utilisation) and about its migration conditions (e.g., minimal battery level or maximal resource utilisation to keep fully/optimally functional provider). According to this information and a particular migration decision strategy, the migration controller finds the most underperforming provider which needs migration of its services (this step will be referred later as "origin provider" decision).

 ISBN: 978-1-61208-280-6

After that, the controller checks every migratable service of the provider selected in the previous paragraph. The migratable services have to publish descriptions of required resources by means of the above defined migration conditions (e.g., the conditions can describe that a particular service can be provided by a provider with at least 180 kB of free RAM memory). According to the information provided by these services of the selected provider and according to a particular migration decision strategy, the controller finds the most appropriate service to be migrated (this step will be referred later as "migrated service" decision).

Finally, the controller starts looking for the best candidate provider which will serve as the migration destination (this step will be referred later as "destination provider" decision). This provider is selected according to a particular migration decision strategy, the state information previously provided by the providers (which describes their available resources), and the information provided by the migrated service (which describes the required resources).

B. Migration Decision Modelling

The migration decision can be described by Linear Time Logic (LTL) [3] as a sequence of states which are related to time. LTL formulae are combinations of terms using logical operators $\wedge$ and $\rightarrow$ and temporal operators $\Box$, $\diamond$, and $\circ$. Formulae $\Box p$ and $\diamond p$ means that p always or sometimes holds in the future, respectively, and $\circ p$ means that p is true in the next state.

Our work is based on [4] which provided a service migration logical framework based on LTL. We focus on the first migration phase and logically describe migration decisions by LTL meanwhile [4] dealt with the rest of the migration phases.

Let $P = \{P_1, P_2, \ldots, P_m\}$ is a set of existing providers and $S = \{S_1, S_2, \ldots, S_n\}$ is a set of the migratable services. Migration decision process D, which has been described informally in the previous section, can be defined as follows:

$$\begin{aligned} D \equiv\ & (D \uparrow \wedge OriginProvider \uparrow \wedge T_a) \\ & \wedge \circ (OriginProvider \downarrow \wedge MigratedService \uparrow \wedge T_b) \\ & \wedge \circ (MigratedService \downarrow \wedge DestinationProvider \uparrow \wedge T_c) \\ & \wedge \circ (DestinationProvider \downarrow \wedge T_d) \end{aligned} \quad (1)$$

In the formula above, $\uparrow$ and $\downarrow$ represent the start event and the end event of each process, respectively, and $T_a \leq T_b \leq T_c \leq T_d$ represent the corresponding events' times.

The migration decision process is described in accordance with the previous section as a composition of three sub-processes, namely: *OriginProvider* to find the most critical provider in the system that needs to migrate its services regarding its current state and its migration conditions; *MigratedService* to find the most appropriate migratable service to be migrated from the "OriginProvider"; and *DestinationProvider* to find the best destination provider for "MigratedService".

The low performance condition of provider P_i will be met sometime in the future when the migration decision process will start. Then, the current provider's performance $ProviderLevel$ will not satisfy preferred performance $DefaultLevel$ of the provider according to a migration decision strategy:

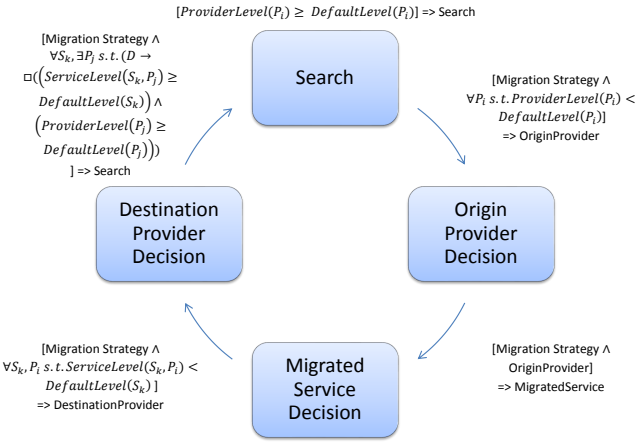


Figure 1. The migration-decision as a finite state automata.

$$\forall P_i \ \ s.t. \ \ (ProviderLevel(P_i) < DefaultLevel(P_i)) \rightarrow \diamond D \quad (2)$$

Sub-process *MigratedService* will be started sometime in the future to determine the most appropriate service to be migrated after a decision about "OriginProvider" has been made:

$$(D \wedge OriginProvider) \rightarrow \diamond MigratedService \quad (3)$$

The necessity condition, the pre-condition of *DestinationProvider* process, becomes true when $ServiceLevel$ describing the current quality of service S_k for $k \in \{1, \ldots n\}$ running on provider P_i for $i \in \{1, \ldots m\}$ is no longer in its preferred level which is denoted as $DefaultLevel$:

$$\begin{aligned} \forall S_k, P_i \ s.t. \ & ((ServiceLevel(S_k, P_i) < DefaultLevel(S_k)) \\ & \wedge OriginProvider) \rightarrow \diamond DestinationProvider \end{aligned} \quad (4)$$

The post-condition of decision process D guarantees acceptable performace of destination provider P_j, where $j \in \{1, \ldots, m\} \setminus \{i\}$, and the necessity condition of its service S_k:

$$\begin{aligned} & \forall S_k, \exists P_j \ \ s.t. \ \ (D \rightarrow \Box(\\ & (ProviderLevel(P_j) \geq DefaultLevel(P_j)) \\ & \wedge (ServiceLevel(S_k, P_j) \geq DefaultLevel(S_k))) \end{aligned} \quad (5)$$

When the migration controller service starts the migration decision process, the output of each sub-process is determined by evaluating provider's and service's conditions. Figure 1 shows the controller's four states ("search", "origin provider decision", "migrated service decision", and "destination provider decision") as a finite state automata with corresponding pre- and post-conditions. The transitions are labelled by $[Conditions] \Rightarrow Process$ where $Conditions$ are the migration conditions of providers and services and $Process$ is a particular sub-process of the migration decision process.

C. Migration of a Serivce

The service migration itself can start when a particular service is selected to be migrated to a particular destination service provider, which is described in the previous section. Then, the migration controller starts migrating the service by

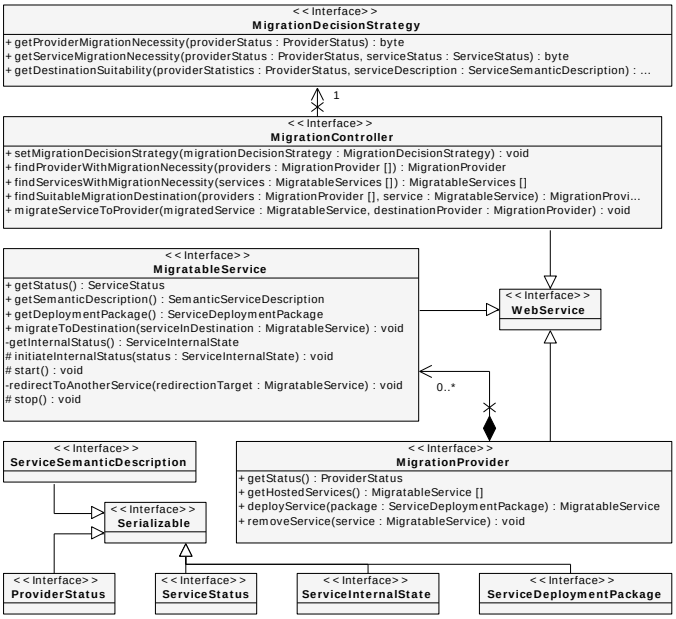


Figure 2. The interface of a control service provided by the framework and the interfaces implemented by participating services and service providers to enable the service migration.

getting a deployment package of the service and deploying it to the destination provider. During this process, the migrating service is stopped and its internal state is stored and sent to the destination provider. All further incomming calls of the service are postponed until the migration is completed, i.e., until the migrated service is initiated in the new location, its internal state is restored, and until the service is able to handle incomming messages.

III. A Framework for Web Service Migration

To support the proposed process of Web service migration, we designed a generic framework, which is introduced in this section. The framework describes an overall service-oriented architecture supporting the service migration and defines interfaces which can be implemented to adapt the framework to a particular Web service implementation technology. It also provides extension points for user-defined migration decision strategies, i.e., the strategies deciding when the migration of a particular service is needed and how it will be performed. The framework's architecture is described in Figure 2 by classes representing specific services with defined interfaces.

To utilise the framework, an implementation of interface *MigrationDecisionStrategy* and auxiliary classes with interfaces *ProviderStatus*, *ServiceStatus*, and *ServiceSemanticDescription*, representing state and semantic information, have to be provided. Migration decisions are based on state information extracted from service providers (e.g., available resources, system workload, battery state, etc.) and their services (e.g., utilised resources, number of requests per an unit of time, etc.) and on the services' semantic descriptions (e.g., provided functionality, inputs and outputs, required runtime conditions, etc.). The migration decision strategy has to be able to acquire instances of the mentioned classes (i.e., the objects representing the state and semantic information) from providers supporting service migration and migratable services.

Interface *MigrationDecisionStrategy* defines methods *getProviderMigrationNecessity*, *getServiceMigrationNecessity*, and *getDestinationSuitability*. The first two methods decide whether some services of a particular service provider or a particular service of this provider, respectively, need to be migrated for some reasons. The third method decides whether a particular provider can be a migration destination for a particular service (i.e., whether a given service can be provided by a given provider after the migration). Returning values of the methods are directly proportional to the necessity of migration of the services or the suitability of the migration destinations.

To be able to migrate, services need to implement interface *MigratableService* with the following public methods. Method *getStatus* returns state information that is used in a migration decision strategy to decide whether a particular service needs to be migrated. Method *getSemanticDescription* provides a semantic description of a migrated service which is used in a migration decision strategy to select a appropriate destination service provider. Method *getDeploymentPackage* returns a service deployment package which is used to deploy a new instance of a migrated service at a destination service provider. Finally, *migrateToDestination* transfers a service's internal state from the service's old instance to its previously deployed new instance and finalises the migration.

Service providers with migratable services need to implement interface *MigrationProvider* with the following public methods. Method *getStatus* returns state information that is used in a migration decision strategy to decide whether services hosted by a particular service provider need to be migrated. Method *getHostedServices* returns all migratable services of a service provider. Method *deployServiceFromPackage* should be able to deploy a service package to create a new instance of the deployed service on a destination service provider. Finally, method *removeService* removes a migrated service from its origin provider.

A. Migration Controller Service

The migration controller is a service provided by the framework which orchestrates services and service providers participating in the service migration. The controller is provided as a "black-box", i.e., it can not be modified and prospective utilisation of the framework can be done solely by implementations of particular migration strategies, migratable services, and their providers. In this way, the controller is also technology independent. It does not need to interpret state information regarding migrated services or their providers and to know how the information has been acquired, it does not need to know a particular technology for service deployment, etc.

The controller implements interface *MigrationController* and orchestrates participating services and service providers as it is described in Figure 3. It periodically checks available service providers for their state information and uses a migration decision strategy to decide whether the information indicate demands for service migration (step (2) in Figure 3). In the case of a service provider which needs service migration, the controller obtains state information from the provider's services (3) and uses the migration decision strategy to decide which

 ISBN: 978-1-61208-280-6

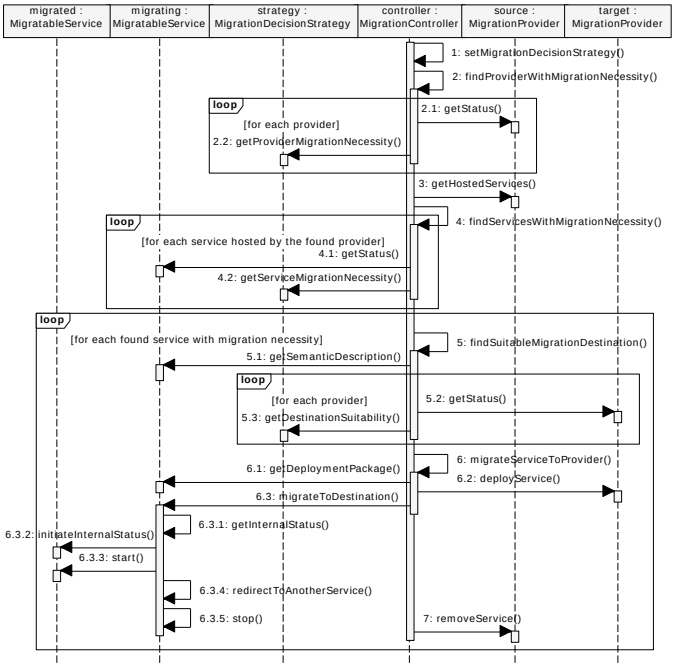


Figure 3. The controller orchestrating service migration.

services need to be migrated and in which order (4). For each such service, the controller obtains its semantic description and tries to find, by means of the migration decision strategy, a service provider suitable as the service's migration destination (5). After that, the controller obtains a deployment package of the migrating service from its origin provider (6.1), deploys a new instance of the migrated service to the destination service provider (6.2), and initiates state transfer from the service's old to its new instance (6.3). Finally, the controller removes the old (inactive) instance from the origin service provider (7).

IV. Related Work

Several works directly address or touch on the migration of components in component-based systems or services in SOA.

Lange et al. [5] described an implementation of mobile agents in Java by the Aglets framework. The framework allows reusing system components, i.e., aglets, in different contexts, however, without any utilisation in making migration decisions.

Hao et al. [6] developed a Web service migration environment and used a genetic algorithm to find the optimal or near-optimal migration decisions. The algorithm calculates the cost of a total round trip including dependency calls for each service request and it is used to decide migration according to this cost. However, the authors did not take into account user-defined conditions affecting the migration decision, e.g., specific requirements on a migrated service or a destination provider.

Zheng and Wu [7] presented an infrastructure for runtime migration of services in a cloud which consists of five components with different roles and of specific criteria to control the migration decision. One of the components collects CPU load data from all known hosts. Then, when the CPU load of a particular host reaches a predefined threshold, a flag is set to indicate that service migration is needed on this host. The approach does not check compatibility of services and providers during migration and does not address possibility of running several services on a single provider at the same time.

Schmidt et al. [8] implemented a prototype of an adaptive Web service migration with two types of migration possibilities, namely context-based migration and functionality-based migration. In the context-based migration, services are migrated to the providers which meet the services' requirements, while in the functionality-based migration, the services are migrated according to their roles in a workflow (i.e., to optimise their communication in the workflow). Both of these migration possibilities can be implemented also in our approach by an appropriate migration decision strategy.

Messig et al. [9] proposed to provide service migration facility in Service Oriented Grid environment which enables taking migration decision based on matching providers' and services' needs and requirements. In this approach, services are hosted by service providers including the resources needed for execution of the services' operations. The authors made several experiments of service migration between two geographically sparse grids where the first grid had high-performance devices and faster network than the second one. While these experiments demonstrated the process of service migration, they are not suitable for the demonstration of migration decisions (e.g., selection of a migration destination) which should be discussed in more detail.

V. Discussion

Contrary to the previously mentioned approaches, the approach proposed in this paper provides a general framework without preference of a particular algorithm or technology for the service migration. The migration decision algorithm and the service migration technology are abstracted by the proposed interfaces (see Section III) and can be implemented and fully customised during utilisation of the framework. The abstraction makes our approach and the corresponding framework more flexible and applicable to different use cases (e.g., in the case of Web services acting as mobile agents), however, it is limited by the proposed architecture and interfaces between the abstracted components and the rest of the framework (e.g., the agent-based Web services may require local migration decisions based on their individual beliefs, desires and intentions, not the centralised approach represented by a single migration decision strategy). Despite the mentioned limitation, our approach is convenient in the cases where the service migration is utilised to keep functionality and quality of services of a service-oriented system and to cope with its inherent environmental dynamics.

Service migration can keep survivability of a system in case of its defect or an attack which damage the system's components or their resources. In this case, endangered services previously hosted by the damaged components can be migrated and started in new and safe locations. Our framework supports this scenario by the migration conditions and the migration decision concepts which allow to detect the damaged components as origins of the migration and the safe locations as possible migration destinations. Focusing on the migration process itself and ignoring the preceding migration decision phase, which is quite common for the approaches mentioned in the previous

 ISBN: 978-1-61208-280-6

section, is not sufficient for the utilisation of service migration in achieving critical system survivability.

Finally, we should discuss a cost of service migration which has not been mentioned before, however, it is important factor of migration decisions as well as of implementations of the migration process. The "migration cost" can be defined as a quantitative metric of difficulty of the service migration. Services in SOA should be designed with respect to SOA principles, and therefore, they should be reusable and stateless [10] which make the service migration easier. Unfortunately, in practice, it is often necessary to break some of these principles, e.g., to use stateful services. In the case of breaking of the SOA principles, migration of the resulting services is more difficult, e.g., it may necessary to migrate resources or state information along with a migrated stateful service. To reflect this issue, the migration conditions and the migration decision proposed in our approach can consider also the migration cost and, for example, migration of stateless services can be preferred to migration of stateful services.

VI. Future Work

The framework proposed in this paper is still a work in progress and needs further elaboration. The future work will focus mainly on implementation of a fully functional prototype of the framework and on its evaluation with particular Web service technologies and migration strategies, covering the scenarios and open issues described in the previous section. Specifically, the future implementation can be divided into two stages following the migration-decision and migration processes.

The migration-decision process which precedes the migration will utilise ontologies and ontology reasoning for description of classes, properties, and relationships, of services, services providers, and devices potentially acting as the service providers. The ontology reasoning will be utilised to evaluate migration conditions, i.e., to nominate services and service providers demanding the service migration, and to select the best destination for migrated services as it is described in Section II-A. We already finished the first version of the ontology and we are currently working on automation of the migration-decision process by ontology reasoning tools.

The implementation of the migration process which performs the actual Web service migration will consist of several steps as it is described in Section III-A. At first, a deployment package and an internal state representation of a migrated service in its original location are obtained. Then, the service is deployed from the package to its new location and the resulting new instance of the service is set to the previously obtained internal state. At this stage, all state-modification operations processed by the service in original location are mirrored to the service in the new location, which keeps internal states of both services synchronous. Finally, the service instance in the new location is activated and the service instance in the original location is deactivated and removed. After that, all incoming messages will be processed by the service in new location, which can be ensured by updating of service registries (i.e., removing the old and adding the new location of the service into a registry) and by forwarding messages going to the original location to the service in the new location (e.g., by a forwarding proxy or by HTTP response code 301 "Moved Permanently" for HTTP-based Web services).

For the service deployment package, native deployment package formats and SOA server management interfaces will be used, e.g., Web Application Resource (WAR) files in the case of Web services implemented in Java EE. The service internal state information will be represented in XML by XML serialisation, e.g., by means of Java API for XML Binding (JAXB) in the case of Web services implemented in Java.

VII. Conclusion

In this paper, we introduced the approach to migration of Web services. We described the process of the service migration including migration decision making and transfer of migrated services to their destination service providers.

Contrary to the previously existing approaches, the approach proposed in this paper provides a general framework without preference of a particular algorithm or technology for the service migration. By introducing fully customisable migration strategies, which evaluate migration conditions and take migration decisions at runtime, our approach can be used to utilise the service migration in different scenarios, e.g., to achieve critical system survivability by the service migration or to assess a cost of the eventual service migration.

Acknowledgment

This work was supported by the research programme MSM 0021630528 "Security-Oriented Research in Information Technology" and by the BUT FIT grant FIT-S-11-2.

References

[1] F. Kon, F. Costa, G. Blair, and R. H. Campbell, "The case for reflective middleware," Commun. ACM, vol. 45, no. 6, Jun. 2002.

[2] A. Michlmayr, F. Rosenberg, C. Platzer, M. Treiber, and S. Dustdar, "Towards recovering the broken SOA triangle: a software engineering perspective," in Proceedings of the 2nd International Workshop on Service Oriented Software Engineering. New York, NY, USA: ACM, 2007, pp. 22–28.

[3] Z. Manna and A. Pnueli, The temporal logic of reactive and concurrent systems. New York, NY, USA: Springer-Verlag New York, Inc., 1992.

[4] Y. Zuo, "Towards a logical framework for migration-based survivability," in Proceedings of the 7th Annual Symposium on Information Assurance / Secure Knowledge Management, Jun. 2012, pp. 29–33.

[5] D. B. Lange and M. Oshima, Programming and deploying Java mobile agents with Aglets. Addison-Wesley, Aug. 1998.

[6] W. Hao, T. Gao, I.-L. Yen, Y. Chen, and R. Paul, "An infrastructure for web services migration for real-time applications," in Second IEEE International Symposium on Service-Oriented System Engineering. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2006, pp. 41–48.

[7] L. Zheng and S. Wu, "An infrastructure for web services migration in clouds," in International Conference on Computer Application and System Modeling. Los Alamitos, CA, USA: IEEE Computer Society, Oct. 2010, pp. 554–556.

[8] H. Schmidt, R. Kapitza, F. J. Hauck, and H. P. Reiser, "Adaptive web service migration," in Proceedings of the 8th IFIP WG 6.1 International Conference on Distributed Applications and Interoperable Systems. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 182–195.

[9] M. Messig and A. Goscinski, "Service migration in autonomic service oriented grids," in Proceedings of the sixth Australasian workshop on Grid computing and e-research, vol. 82. Darlinghurst, Australia: Australian Computer Society, 2008, pp. 45–54.

[10] T. Erl, Service-Oriented Architecture: Concepts, Technology, and Design. Upper Saddle River, NJ, USA: Prentice Hall PTR, Aug. 2005.

 ISBN: 978-1-61208-280-6

Web Service and Structure of University Data

- Development of Japanese Higher Education Database -

Masaaki Ida
Department of Research and Development
National Institution for Academic Degrees and University Evaluation
Tokyo, Japan
ida@niad.ac.jp

***Abstract*—This paper describes the state of research and development of web services and data structure of university survey data, especially, Japanese higher education survey data. Our research and development are aimed for university comparative analysis on the data with consideration on general university information structure. Institutional data of university, college, or college of technology are substantially important for data analysis or knowledge discovery in the higher education management field. However, university institutional data are not necessarily standardized and compiled, so it is not easy to integrate their information for various reporting and data analysis. In the past decade, investigation of the integrated university data sets has been done to deal with various kinds of university institutional information including university survey or school basic survey data in Japan, by providing the structured university data via web services. This paper describes the state of research and development of web services and data structure of university survey data, which are utilized for understanding the university characteristics. We describe: (i) a proposal of the XML schema for Japanese university data (ii) development of various Web APIs (XML, JSON) of university database for survey cards.**

Keywords- database; school basic survey; web service; data structure.

I. INTRODUCTION

A. University Insititutional Data

Development of education-related databases is substantially important for data analysis and knowledge discovery in education field worldwide [1]. Institutional data of universities are not easy to analyze because they are not necessarily standardized and integrated in each university itself or national education-related agencies. However, some advanced education-related database systems are progressively developing.

In the United States, *Integrated Postsecondary Education Data System* (*IPEDS*) [2] of National Center for Education Statistics has been developed to collect and analyze basic institutional information about universities and colleges in U.S. IPEDS standardizes and accumulates the information nationwide. This system comprehensively holds basic institutional data, such as institutional characteristic, degree completion, enrollment, human resource, finance, student financial aid, graduation rate, and so on. Moreover, this system is equipped with facilitated data analysis tools to conduct university comparative analysis.

College Portrait [3] is also higher education database that is "a source of basic, comparable information about public colleges and institutions presented in a user-friendly format".

In European area, for example, *Unistat* [4] system is developed in order to search, review, and compare subjects at UK universities and colleges. It "is the official website to help you make an informed choice when deciding which UK university or college to apply to. It includes the results of the latest National Student Survey".

In Asia, Korean government started their university evaluation system which consists of the university information disclosure system, *Korea Academyinfo* [5], conducted by the Korean Council for University Education, so that higher educational information of Korean universities is published on their web site.

These databases are well-organized and comprehensive systems with easy web-based operation on their web pages. However, in order to cooperate (mash-up) with other information systems, e.g., in-house database developed in individual higher education institutions, or external web services (Google Chart API), more improved systems should be equipped with various web service functions and standardized data sets.

B. University Basic Survey in Japan

In Japan, Ministry of Education, Culture, Sports, Science and Technology collects basic information about higher education institutions in Japan [6]. This basic survey data include the yearly information of higher education institutions, such as number of faculties or staffs, number of enrolled students by grade (undergraduate, graduate, foreign student), number of graduates by subsequent course, number of those who were employed after graduation by industry and by occupation, faculties, facilities, and financial data. Fig. 1 shows an example of university basic survey sheets.

Its online submission system via internet has been developed, which is equipped with authentication and encryption functions. Persons of universities in charge of data

 ISBN: 978-1-61208-280-6

submission must fill the data into electronic files (PDF files) and submit them through internet to the ministry's data collection server.

The data in the files are saved as XML data as duplication data of the submitted PDF file data. However, the XML data structure is designed only for data submission purpose. And it is not designed for data analysis purpose. These submitted survey data are compiled and published by the ministry as summarized statistics data tables. Parts of the survey data of some universities are published on their web sites. However, under the present circumstances sufficient amounts of data, detailed and standardized data required to conduct intercollegiate comparative analysis are not necessarily obtained. It is difficult to examine detailed situation of higher education institutions from various perspectives.

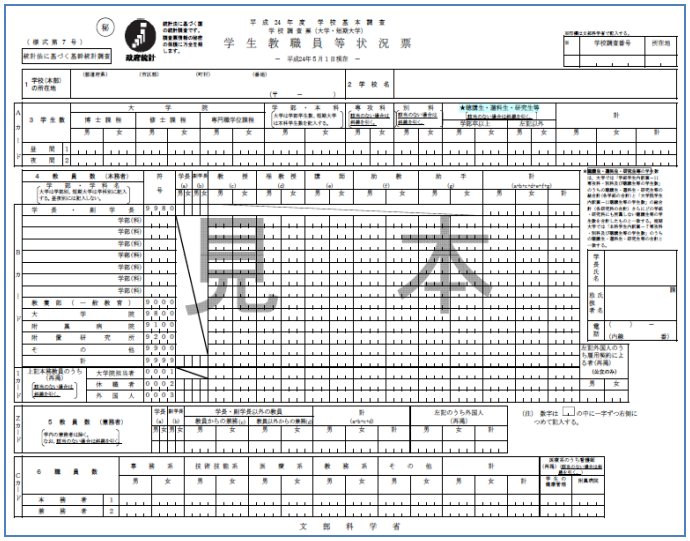


Fig. 1. Survey sheet (number of faculty and student) [7].

C. eXtensible Business Reporting Language

The *eXtensible Business Reporting Language* (*XBRL*) [8-10] is one of the computer languages based on XML, which is a standard for the electronic exchange of data between businesses on the internet. XBRL utilizes some XML technologies such as XML Schema and XLink standards. Based on XML, tags are applied to items of financial business data so that financial data can be processed efficiently by computer software. XBRL is implemented in a wide range of scenes such as tax payment system and financial data transfer system in stock exchange.

XBRL consists of an *XBRL Instance*, containing primarily the business facts being reported, and a set of *Taxonomies*, defining metadata about these facts, such as what the facts mean and how they relate to one another:

- *Instance* holds the following information: business facts, contexts (date and time information, scenario), units, footnote, and references.
- *Taxonomies* are the reporting-area specific hierarchical dictionaries. The XBRL specification defines five different kinds of linkbases (*Label* linkbase, *Reference* linkbase, *Definition* linkbase, *Calculation* linkbase, and *Presentation* linkbase). Taxonomies consist of hierarchical structure.

Different taxonomies are required for different purposes in various application fields, therefore, we extend or modify taxonomies for university information.

In the following sections of this paper, Section II presents structured data and university information database system. Section III presents construction of university survey database and schema of university survey data. Section IV presents web service and application of university survey data.

II. UNIVERSITY INFORMATION DATABASE SYSTEM

A. Structured Data Sets and Taxonomies

In this section, we describe the general information structure of university information and the XBRL extension for university information as shown in Fig. 2.

As shown in the lower right side of the Fig. 2, various data and databases are developed in each field such as enrollment, finance, personnel data, and so on, which possess and provide their data in various manners cooperating with other databases. These databases provide specified and designated information as HTML files or XML files via Web API [11].

In the upper left side of the figure, taxonomy hierarchical structure expresses the generalized information structure with specified taxonomy levels, three level structure, such as national or international standard level, institution type level, and individual institution level. The structured taxonomy sets and their data stored in various concrete databases are linked mutually, so that rigid and various definitions of data and hierarchical structures are possible, and flexibility for the changes of database equipment and time transition of data definition is guaranteed.

Their data are transmitted via web services, which are composed of REST type Web APIs (JSONP) with XML and JSON data transferred into the Business Intelligence system as shown in the right-hand side of the figure. The Business Intelligence system produces integrated university reports and the results of comparative analysis of higher education institutions combined with outer databases, reporting system, and analysis systems (mash up).

 ISBN: 978-1-61208-280-6

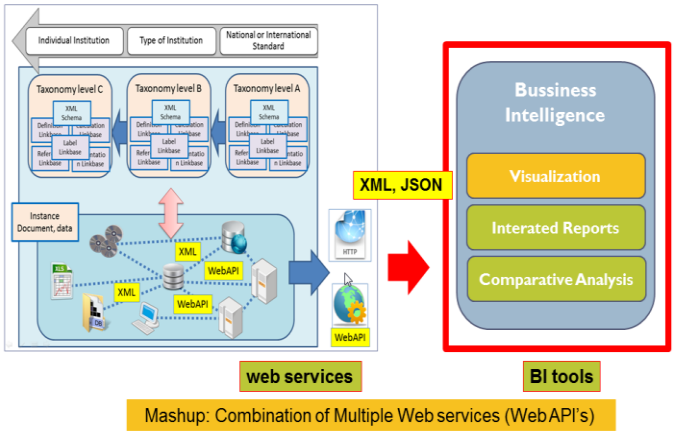


Fig. 2. University information system and taxonomy hierarchical structure.

B. *Extension of Taxonomies*

We conducted taxonomy design for university financial statements in Japan [12], i.e., we proposed taxonomy for *Balance Sheet* and *Income Statement* of university, which is an extension from ordinal XBRL taxonomy. The structure of proposed taxonomy for university financial statement consists of five different kinds of linkbases or files. In the taxonomy, the definitions of the vocabulary (Japanese and English) are assembled in the definition file folder for general company or in the extended university finance file folder. The presentation orders or format structure of the financial elements in the financial statement are defined in the university finance presentation file folder.

Adding to the extension for university finance data, we proposed further extension for university institution data [13], which aims for expressing various indexes of institutional situation, e.g.,

Ratio of education expenses to student
= (*Education expenses*) / (*Number of student*).

In order to treat various indexes related to higher education institutions for institutional reports, we extended the general hierarchy of terminologies, e.g., taxonomy extension such as *number of undergraduate male* and *under graduate female* extended to ordinal taxonomy.

III. UNIVERSITY SURVEY DATA

A. *Construction of University Survey Database*

In this section, we focus on the data of *University Basic Survey* and its database. We have examined the data structure of university survey of Japan. Based on the examination we are considering database system as an infrastructure for applying them to analyze various aspects of universities. As a pilot system, we tried to collect sample survey data of U.S. and store them into the developed database so far.

Up to now for Japanese university survey data, we designed and developed useful tools that upload and transform the XML data in university basic survey, and store them into a relational database. The data flow and process for the survey are shown in the Fig. 3.

University Survey Formats (sheets) are consist of detailed university information card, such as "Institutional structure, faculty member and staff (7_A (number of students), 7_B1 (number of academic staffs), 7_C, 7_Z)", "Student (8_D2 (number of students of each department), 8_3, 8_E, 8_G, 8_7, 8_R)", "Graduate student (9_H4, 9_5, 9_I, 9_S, 9_8)", "College (10_J6, 10_9, 10_K)", "Foreign student (11)", "Facility (20)", "Financial data (22A, 22B)", "Employment (30)", and so on. The detailed card information, such as card 7_A, card 22A, are described in reference [7].

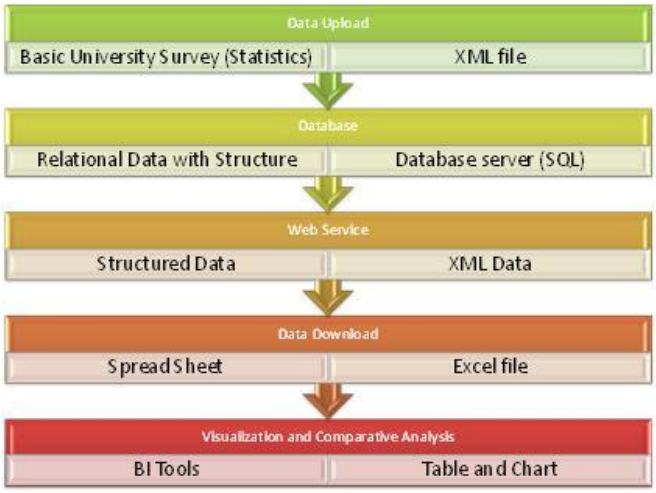


Fig. 3. Data flow and process for Japanese university basic survey.

B. *Schema of University Survey Data*

In this paper, we propose XML schema of various university information in University Survey Formats. The XML schema represents the University Survey information of several universities into a standard and integrated university formats.

Fig. 4 shows the overview of XML Schema structure of university information. Each dashed-green rectangle indicates university level schema, and dashed-blue rectangle indicates department level schema. Fig. 4 is a part of whole structure of the XML Schema of university survey information.

 ISBN: 978-1-61208-280-6

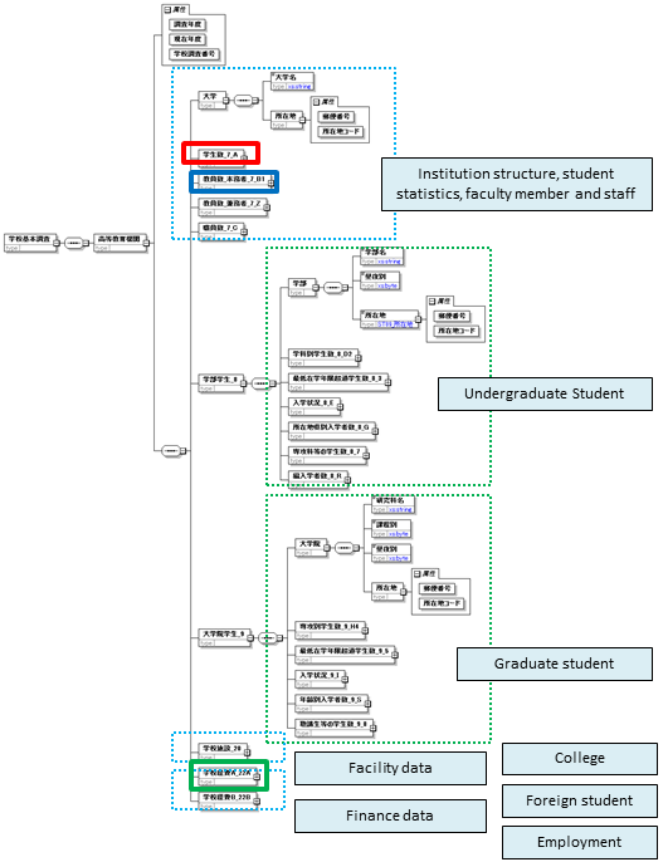


Fig. 4. XML Schema for Japanese university basic survey.

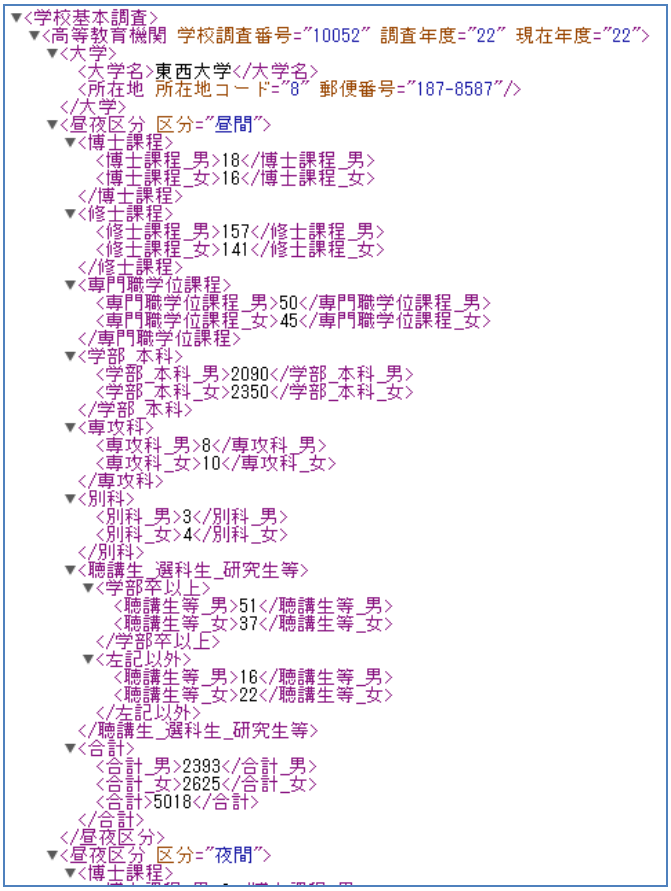


Fig. 5. Output example of web service (Number of student; Format 7).

IV. WEB SERVICE OF UNIVERSITY SURVEY DATA

A. *Web API of University Survey Database*

In this paper, we develop two kinds of Web APIs (XML, JSON) of university database for survey cards.

Several web APIs were considered which are suitable for data analysis and data dissemination. This type of web services cause independency of application modules which can be easily redesigned and reformed.

The following are examples of RESTful web service (Web API) [14] retrieved by survey year and institution number of Japanese universities and so on.

Fig. 5 shows an example of output of web service on the number of student form the university survey (survey format 7), that is corresponding to red rectangle part in Fig. 4. Elements in Japanese mean “university name”, “address”, “number of undergraduate student”, “number of graduate student (master)”, “number of graduate student (doctor)”, and so on.

Fig. 6 shows an example of output of the web service concerning university financial data from the university survey (survey format 22) , that is corresponding to green rectangle part in Fig. 4. Elements in Japanese mean “university name”, “address”, “faculty salary”, “staff salary”, “education expense”, “management expense”, and so on (unit: 1,000 Yen).

Fig. 6. Output example of web service (Finance; Format 22).

B. Web API and Reporting

Receiving the web service data form database system at the client side, this system generates spreadsheet or PDF files, which are *reporting sheets* subject to the conventional formats of university survey booklets.

When the XML data from web services are obtained, that is corresponding to blue rectangle part in Fig. 4. The data are simply transmitted into spreadsheet files with data bindings as shown in Fig. 7. The elements in Japanese mean the number of "professor male", "professor female", "associate professor male", "associate professor female", and so on for each university faculty. The Excel sheets only possess the relationships between items of XML returned by web service and the columns of sheets. Therefore, in case that some data on University Database are modified, we don't need to adjust the spreadsheet structure, and the data in each sheet would be automatically changed.

University DB system –> Web API (REST) –> XML –> Data binding –> Excel Sheets

大学	学部	教授		准教授		講師		助教	
		男	女	男	女	男	女	男	
Name	FacultyNan	ProfessorM	ProfessorF	AssociateF	AssociateF	LecturerMa	LecturerFe	AssistantF	Assis
A医科大学	医学部	20	50	40	20	40	20	40	
A医科大学	歯学部	34	53	45	60	56	57	58	
A医科大学	看護学部	62	63	64	32	66	67	68	
A医科大学	薬学部	15	15	15	15	15	15	15	
B大学	文学部	46	9	26	11	8	1	1	
B大学	教育学部	21	5	32	10	11	8	1	
B大学	法学部	46	42	22	3	10	30	30	
B大学	経済学部	32	45	52	3	23	30	30	
B大学	理学部	29	23	23	23	0	30	30	
B大学	工学部	43	45	10	10	43	52	30	
B大学	生物理工学部	30	30	30	30	11	9	26	
B大学	農学部	30	30	30	30	23	5	32	
B大学	環境学部	32	30	30	30	0	30	22	
B大学	経済学部	62	52	20	50	40	20	1	
C工業大学	理学部	46	9	34	53	45	60	66	
C工業大学	工学部	20	5	62	63	64	32	15	
C工業大学	理工学部	32	30	15	15	15	15	8	
C工業大学	システム工学部	10	65	46	9	20	50	40	
C工業大学	生物理工学部	64	32	21	5	34	53	45	
C工業大学	法経学部	15	15	46	42	62	63	20	
D大学	文学部	26	11	32	45	15	15	34	
D大学	教育学部	32	10	29	20	50	40	20	
D大学	法学部	22	3	43	34	53	45	60	
D大学	経済学部	52	3	23	62	63	64	32	
D大学	理学部	23	23	0	15	15	15	15	
D大学	工学部	10	10	43	46	9	26	11	
D大学	農学部	30	30	11	21	5	32	10	
D大学	医学部	30	30	23	46	42	22	3	
D大学	薬学部	30	30	0	32	45	52	3	
D大学	国際教養学部	30	40	20	29	23	23	23	

Fig. 7. Reporting results: Binding of spreadsheet and web service.

C. Web API and Comparative Analysis

We can utilize these web services in case of comparative analysis. In this section, we show some examples of comparative analysis using JSON data derived from the web services and mash up with outer web APIs. Simplified and smart analyses can be executed as shown in the following examples. Outer web APIs are included in Google API (Google Chart, Map, …).

DB system –> Web API (JSONP) –> JSON –> "Mash Up" -> JavaScript (jQuery) –> Visualization (Chart, Graph)

1) Comparative Analysis on Ration of Expenditure

Fig. 8 shows an example of comparative analysis of three major national universities on expenditure ratio. The data from the web API is JSON data of JSONP call back function, so that this analysis is programmed by JavaScript combined with "*Google Chart API*". Mash-up programing is not difficult for beginner of programming because of the typical combination of web services for university comparison and analysis.

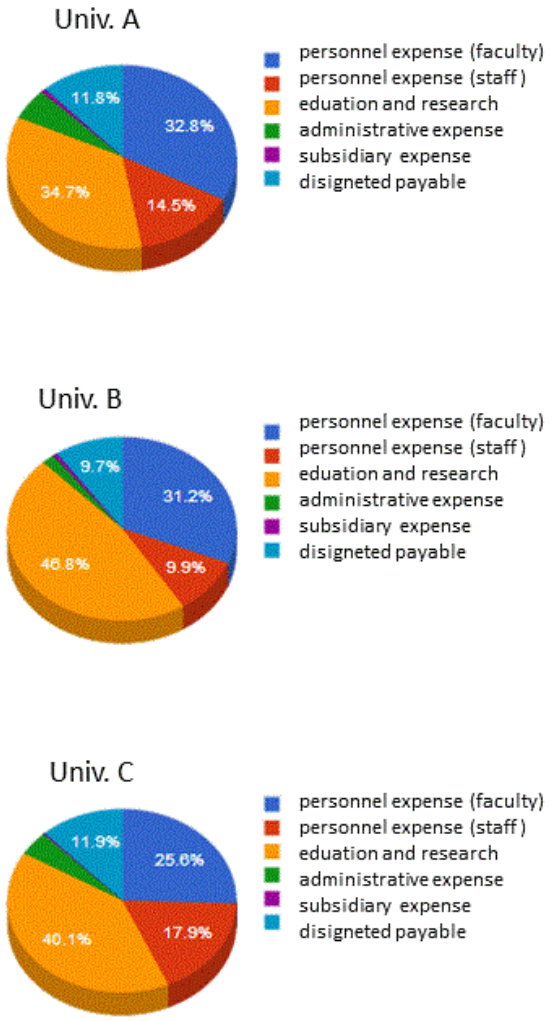


Fig. 8. Example of university survey data analysis: Annual expenditure of three national universities in Japan.

2) Comparative Analysis on Faculty Salary

Fig. 9 shows an example of university survey analysis on total faculty salary on various sections (faculty, university hospital, research institution, total amount) on four major national universities in Japan. This visualization is programmed by JavaScript combined with some JavaScript libraries such as jQuery.

 ISBN: 978-1-61208-280-6

V. CONCLUSION

Applications of education related information are substantially important for data analysis and knowledge discovery in education field. This paper described the state of research for web service and data structure of university survey data, which is utilized for analyses of university characteristics. In this paper, we (i) proposed the XML schema for Japanese university data, and (ii) developed various Web APIs (XML, JSON) of university database for survey cards. In order to handle of more general university data such as the data between some countries, we have to coordinate differences between those data for effective comparisons. We hope that our proposal will play an important role as an infrastructure for data analysis and knowledge discovery in higher education field.

REFERENCES

[1] C. Romero, S. Ventura, M. Pechenizkiy and R. Baker (eds.), Handbook of Educational Data Mining, CRC Press, 2010.

[2] Integrated Postsecondary Education Data System, IPEDS, http://nces.ed.gov/ipeds/

[3] College Portrait, http://www.collegeportraits.org/

[4] Unistat, http://unistats.direct.gov.uk/

[5] Korea Academyinfo, http://www.academyinfo.go.kr/

[6] Ministry of Education, Culture, Sports, Science and Technology, basic information of higher education institutions in Japan, http://www.mext.go.jp/b_menu/toukei/

[7] Example of survey sheet (sheet of "faculty and student"), http://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2012/03/30/1318957_3.pdf

[8] eXtensible Business Reporting Language, http://www.xbrl.org/

[9] R. Debreceny, C. Felden, B. Ochocki, M. Piechocki, et al., XBRL for Interactive Data: Engineering the Information Value Chain, Springer, 2009.

[10] C.Hoffmann, and L.A. Watson, XBRL, Wiley, 2010.

[11] M. Ida, "Web Service and Visualization for Higher Education Information Providing Service," Proc. of ICSESS2010, pp.415—418, 2010.

[12] M. Ida, "XBRL Extension for Knowledge Discovery in Higher Education," the 8th International Conference on Fuzzy Systems and Knowledge Discovery, pp.2177—2180, 2011.

[13] M. Ida, "XBRL Financial Database for Higher Education Institutions," the 14th International Conference on Advanced Communication Technology, pp. 398—401, 2012.

[14] L. Richardson and S. Ruby, Restful Web Services: Oreilly & Associates Inc, 2007.

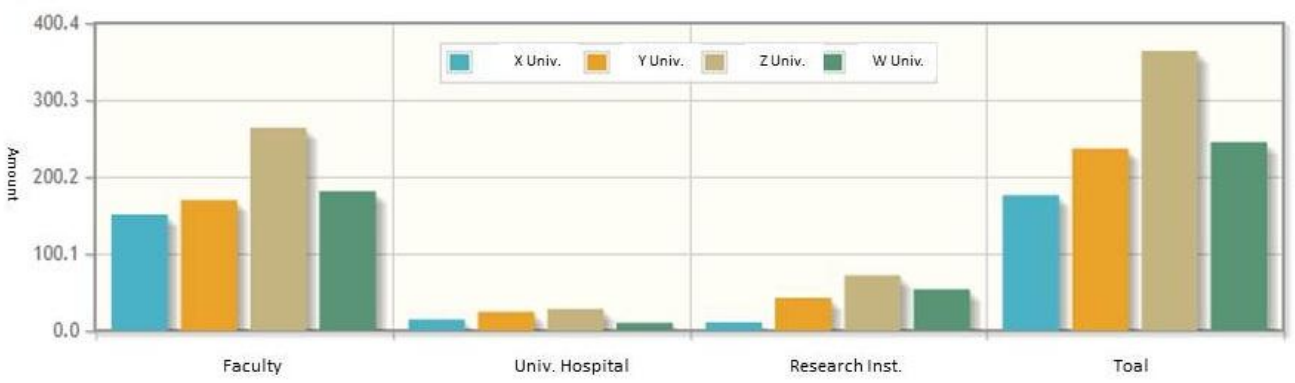


Fig. 9. Example of university survey analysis for financial information: Total salary on sections (10^9 Yen)

A Visual Semantic Search Framework for Finding Craft Services

Maximilien Kintz and Andrea Horch
Institute for Human Factors and Technology Management (IAT)
University of Stuttgart
Stuttgart, Germany
{maximilien.kintz, andrea.horch}@iat.uni-stuttgart.de

Abstract— **When a house is damaged, the house owner needs to quickly seek help from specialized craftsmen. Unfortunately, the complex structure of craft services and the lack of specialized knowledge make this task more difficult than it should be. To solve this problem, a standardized categorization of craftsmen and craft services has been developed in form of the German Crafts and Craftsmen Ontology (GeCCO). In this paper, an easy-to-use web-based tool for browsing and searching the ontology while taking advantage of all its semantic features is presented. Furthermore, a method for combining the ontology with a standard geographical web search tool is presented.**

Keywords-ontologies; graph visualization; Web search; craft services.

I. INTRODUCTION

Finding the nearest craftsman best suited to help you with damage in your house is, even in the time of quick and easy map-based Web searches, harder than it should be: domain specific knowledge such as "who can repair which damage?" or "how are crafts organized?" is often required from house owners, who usually are not craft services specialists at all.

To help users accomplish this task and in order to provide a structured way to find the appropriate service, we propose a visual semantic Web search framework based on an extensive self-created ontology: GeCCO - German Crafts and Craftsmen Ontology. The ontology is presented in detail in a paper published in 2012 (see [1] for more information).

The remainder of this paper is organized as follows: In Section 2, we discuss related work. In Section 3, we present requirements for a semantic search framework and assess the existing tools described in Section 2. In Section 4, we introduce a dynamic graphical user interface for browsing the craft services ontology. In Section 5, we describe the semantic Web search tool to find craft services. Finally, in the concluding Section 6, we present our methodology for evaluating the search tool, discuss limitations of the current implementation and propose ways to further develop the tool and possible future outcomes.

II. RELATED WORK

Related work and state of the art solutions are presented for ontology and graph visualization, semantic search and ontology search solutions, and finally for the description of technical and non-technical services.

A. Graphical Interface for Ontology and Graph Visualization

Many approaches to ontology visualization have been investigated and are today in use. Ontology viewers often already exist as part of ontology editors (examples of well-known visualization plug-ins for ontology editors being OntoGraf [4] for Protégé [5] or NavigOWL [6] for OntoStudio [7]). Katifori et al. have extensively investigated tools for ontology visualization [3]. More generally, ontologies can be viewed as graphs in which nodes represent classes and edges represent relations. Visualization techniques for large graphs have been investigated, e.g. in [8].

However, these tools tend to be targeted at specialists who understand the concepts of ontologies. In contrast, Web users seeking to achieve a simple common task, such as finding the appropriate craft service after damage, are often overstrained by the complexity of generic ontology viewers.

In a previous article, we introduced a tool for the visualization of a specific typo of graph: company business relations [15]. The software technologies used for the implementation of the graphical interface in the present work base on and largely extend this previous work.

B. Semantic Search

Semantic or ontology-based search engines have also already been widely investigated, e.g. by Lei et al. [17] who introduced SemSearch, a search engine allowing users without specific ontology knowledge to benefit from semantic Web technologies for their search queries. Natural language interfaces for semantic search engines have generally widely been investigated, another example being the ORAKEL interface [18]. Our work differs from these approaches in that it introduces a graphical component for ontology browsing and focuses on simple questions for the specific domain of craft services, such as "who can repair a

 ISBN: 978-1-61208-280-6

specific damage". The tool presented in this paper does not require any specific knowledge from the user, neither in semantic Web or ontologies nor in the domain of craft services.

C. Service Description and Matchmaking

A data interchange format is needed for the communication between the search engine in the background and the visualization interface. For this purpose, we use the Unified Service Description Language (USDL) introduced in [9]. USDL offers all required characteristics of service description (such as name and contact detail, capabilities or pricing models) and is better suited to describe generic real-world services, such as in our example craft services, than the possibly better known Web Service Description Language (WSDL) [2], tailored for the description of technical Web services. The limitations of WSDL for finding appropriate services and a solution for semantic service matchmaking were already investigated and presented by Paolucci et al. in [27].

III. Requirements for the Semantic Search Framework for Craft Services

Our goal is to provide users without any specific knowledge in semantic Web or craft services with a powerful user interface to easily find the appropriate craft service company that could help them with their damage. Furthermore, the results should be easily integrated in other software solutions, such as online service marketplaces (R5 in TABLE I). From this given general requirements, we derived five more specific requirements listed in TABLE I. The table also indicates how existing tools and frameworks discussed in the previous section fulfill these requirements.

TABLE I. Requirements for the search framework and existing solutions

Requirement	Existing solution			
	Grapjhical interfaces	***Sem-Search***	***ORAKEL***	***Paolucci et al.***
R1. No need for semantic web knowledge		X	X	
R2. No need for domain-specific knowledge		(X)	X	
R3. Navigation across domain classes and relations	X			
R4. High precision of search results	X	X	X	X
R5. Possibility to integrate in larger framework				X

Although existing solutions cover part of our requirements, a mash-up approach had to be followed to be able to cover all of them at once. In particular, it can be noted that many already existing frameworks provide a high precision of search results, which is a basic requirements for any specialized search tool (R4). However, in our use case and considering our target user-group (people searching for craftsmen and craft services), other requirements are more important. For example, navigation across classes and relations (to discover craft services related to other services or to specific objects, R3) or intuitive entry of query terms and reading of search results (implying that no need of specific knowledge in the specific domain of craft services (R2), or more generally of semantic Web (R1) is needed) were major design criteria. The following sections present the hybrid approach we used to try and cover all requirements.

IV. GeCCO Ontology Browser

In this section, we briefly describe the German Crafts and Craftsmen Ontology (GeCCO) and introduce a tool for visually browsing and searching the ontology to find the appropriate craft services.

A. The GeCCO Craft Services Ontology

Current standardized crafts categorization systems (such as the German Crafts Code or *Handwerksordnung* [10]), do neither include all needed logical connections between crafts and craftsmen nor the synonyms for a sufficient use in a search system. Current craftsmen search systems (such as MyHammer [11]) have their own categorizations of crafts and craftsmen, with the same limitations as the standardized systems.

GeCCO (German Crafts & Craftsmen Ontology) is a new ontology developed as part of the research project openXchange (information about the project can be found online [12] and in print [13]) and introduced in [1]. GeCCO creates an extensible class hierarchy and defines synonyms and all needed logical connections between the classes.

B. Browsing the Ontology with a Dynamic User Interface

Potential users of a craft services search tool do not necessarily have specific knowledge in the classification of craft services. As was already explained in [1], the goal of the GeCCO ontology was to help users navigate through craft services and related objects. Thus, users of a craft services search tools need to be offered a dedicated simple browsing and searching interface.

The Prefuse Flare [14] Flex framework was chosen, as its extended configurability and high performance had given good results in previous own work concentrating on the visualization of company relations [15]. However, to meet the specific needs of the craft services ontology visualization, a tool supporting more advanced functionalities needed to be implemented.

The first important improvement is the integration of a custom built OWL [26] parser that reads the ontology structure as stored in an OWL XML file and creates the graph structure needed by the Prefuse framework. We couldn't identify any already existing OWL parser for use with the Flex technology, which led to the implementation of

 ISBN: 978-1-61208-280-6

this custom-built parser capable of correctly extracting classes and relations from an OWL file describing an ontology.

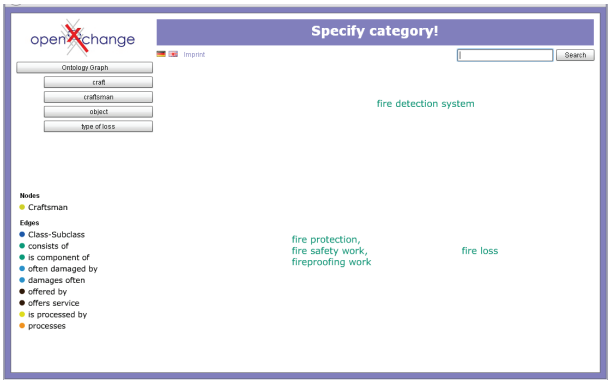


Figure 1. Example search results for a search for "fire"

The tool supports all languages defined in the ontology (in the present case, German and English) by taking advantage of the differentiation introduced in the ontology between unique class names and possibly multiple and language-dependent labels. Synonyms are visually presented, since it is typical for craft services to be known under different names in different regions of Germany (a typical example being Flaschner and Klempner, both referring to a plumber). Furthermore, craft services which used to be distinct services but have merged over the years are also indicated as such.

The browsing starts with an overview of the ontology. By zooming, clicking and panning, the user can navigate to a specific part of the ontology.

Using the search textbox provided on the upper right side, the user can perform a simple text-based search. In this case, the search tool queries the ontology and returns appropriate results as follow:

- if the user entered a word that does not match to a unique class, a circular presentation of all matching classes is returned (Figure 1), letting the user chose the appropriate one or enter a new search query,
- if the user entered the name of a known ontology class, the detailed view of this class is shown (Figure 2).

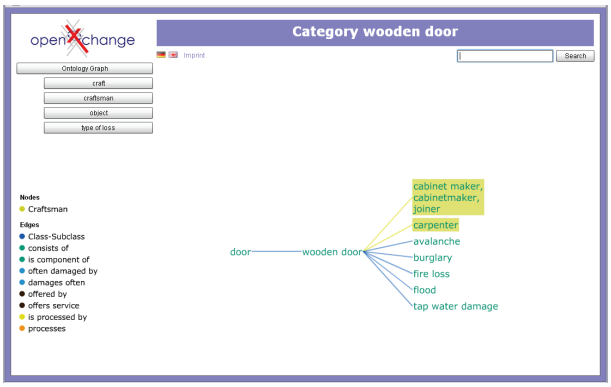


Figure 2. Detailed view of the category "wooden door"

The detailed view of a category of the ontology indicates all related upper-classes and sub-classes, linked by edges colored according to the relation they represent. The key for the color-coding is presented on the left side of the browser window. For example, on Figure 2, blue edges indicate the "often damaged by" relation as well as its inverse "damages often". Classes are represented with their labels. To allow a faster and simpler navigation and differentiation from damaged objects, classes representing craft services are highlighted.

By double-clicking on the name of a craft service, the user can choose between showing the graph for this service or being redirected to a Web-based craft service search tool (see Section 5), so as to find the contact data of the nearest office.

V. CRAFT SERVICES WEB SEARCH TOOL

The GeCCO ontology browser helps users identify the appropriate craft service needed to help them. This identification is the starting point for searching a specific craft service company, for example the one that is located nearest to the user. To that end, a dedicated Web-based search interface for a semantic Web service was built and linked to the ontology browser.

A. *Graphical Interface for Ontology and Graph Visualization*

The craft services Web search front end consists in a single text box, in which a user can enter his search query.

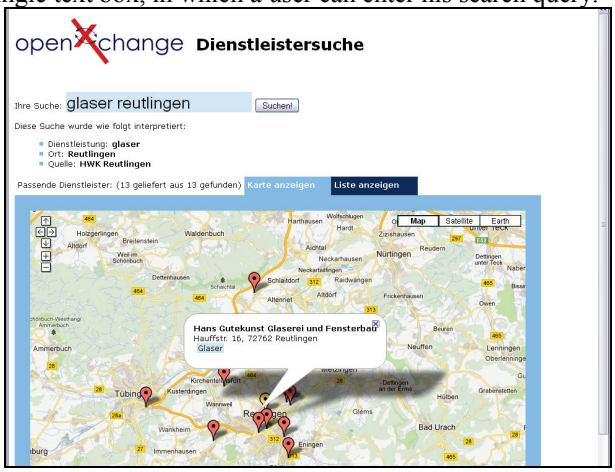


Figure 3. Map-based search result presentation for a search for Glaser (glazier) near the city of Retulingen

The user does not have to use a complex search form to perform advanced search queries: The tool recognizes which parts of the search queries need to be matched with which possible search criteria. Classes defined in the ontology (crafts and goods), using GeCCO as a reference, as well as German cities, using a geocoding service as reference (in the current implementation case, the Google Maps geocoding API was used) can automatically be recognized. Thus, a search for a glazier near the city of Reutlingen (cf. Figure 3) can be entered as "glazier reutlingen" and, using semantic

 ISBN: 978-1-61208-280-6

search capabilities, will be correctly interpreted by the tool as the combination {glazier: craft service, Reutlingen: city}.

The results are retrieved from the Web service and presented to the user in two possible manners.

The first presentation is a Map-based one (also shown in Figure 3). The reference point for the search is indicated by a yellow bubble, the matching craft services by red bubbles. By clicking on a bubble exposes detailed information on the respective company, such as name, available services and contact information. This view is best suited for geographic searching and selection of the nearest service provider.

Figure 4. List-based search result presentation for a search for a Schreiner (carpenter) named Beck

The second presentation is a list-based view of the results (see Figure 4). The same information is presented as a conventional text list allowing for rapid scanning and exporting matching data lists.

B. *Search Tool Architecture*

The architecture of the craft services visual search framework (see Figure 5) can be categorized in three main parts:

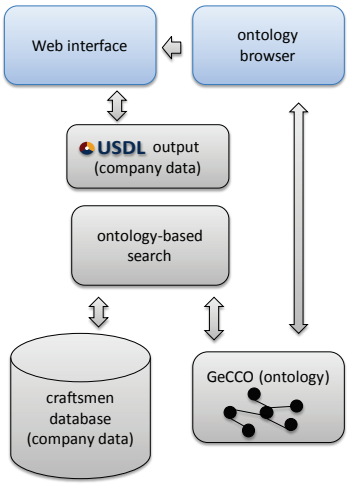


Figure 5. Architecture of the visual semantic craft services search framework

- a repository consisting in the ontology as an OWL XML file and a MySQL [24] database containing craft services company data (addresses, names, etc.),
- a search SOAP Web service, and
- a user interface (presented in sections 4.B and 5.A) consisting of the ontology browser and of the Web-based search tool.

The used technologies are established open-source software components.

```
<?xml version="1.0" encoding="ASCII"?>
<usdl3:ServiceDescription xmi:version="2.0"
xmlns:xmi="http://www.omg.org/XMI"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:foundation="http://internet-ofservices.
com/usdl/foundation/20100416-M4"
xmlns:participantmodule="http://internet-ofservices.
com/usdl/modules/participants/20100416-M4"
xmlns:pricingmodule="http://internet-ofservices.
com/usdl/modules/pricing/20100416-M4"
xmlns:servicelevelmodule="http://internet-ofservices.
com/usdl/modules/servicelevel/20100416-M4"
xmlns:usdl3="http://internet-ofservices.
com/usdl/20100416-M4">
<Guid>d11c40be-5df9-4a02-8e32-ea78323c0f55</Guid>
<Service xmi:id="ServiceWrapper_8732241">
<ServiceElements xmi:id="Service_29245520">
<Guid>bdcb0674-9d90-479f-8fb7-92b6e6142fe0</Guid>
<Version>1.0</Version>
<Name>Name Handwerksunternehmen</Name>
<PublicationTime xsi:type="foundation:AbsolutePointInTime"
xmi:id="AbsolutePointInTime_32657640">
<TimeZone>Europe/Berlin</TimeZone>
<Value>2010-09-17T10:57:25.616+0200</Value>
</PublicationTime>
<Nature>Human</Nature>
<Certifications>
…
```

Figure 6. Excerpt from the XML serialization of the USDL description of a generic craft service

The search Web-front-end consists of a simple Java servlet-based Web application running on an Apache Tomcat [25] server. The Web application is entirely focused on user interface generation and only passes queries to the Web service and presents results in HTML. No interpretation or search happens at that level: all search related activities are performed by the Web service.

The Web application and Web service exchange queries and results using a simple self-specified XML format that is encapsulated in SOAP messages. In the search results messages, the actual service descriptions (i.e. name, address, associated crafts and related information) are encoded using the XML serialization of USDL (see Figure 6. for an example).

The use of the self-defined XML language is restricted to metadata such as the total number of results found, or the way in which the search query was interpreted. USDL was used since it provides a complete description language for services, from basic information such as contact data, up to more complex descriptions of capabilities or pricing models for these capabilities. Thus, the nature of the information being transmitted by the service could be vastly enriched without changing anything to the search service architecture.

To facilitate the integration of the search service in an online service marketplace regrouping other similar or complementary services, the interface, search parameters, and result formats were also described using the USDL language.

 ISBN: 978-1-61208-280-6

The ontology-based search service uses a simple architecture. For the development of the SOAP interface, the Apache Axis2 [23] framework was used, as it allows a very simple integration with standard Java applications. The search engine interprets the search query (i.e. if the ingoing query was "glazier Reutlingen", in the interpreted query "Reutlingen" is mapped to "city" and "glazier" to "craftsman) to build an SQL query and retrieve the results.

The algorithm used to interpret the query works as follows:

1) The query is split in individual tokens
2) Each token is mapped to an ontology class, i.e. a possible craft service category, damaged object, etc. If a token could successfully be mapped to an ontology class, it is removed from the list. If necessary, stemming of fuzzy mapping of keywords and class could be implemented (this was considered not useful for our test scenarios).
3) Non-matched tokens are sent to the Google Maps API [22] geocoding service. If they could successfully be mapped to German cities, they are considered as geographical restriction and removed from the list.
4) Remaining tokens are considered as possible names of craft services providers.

This algorithm gave satisfying results with all basic test examples, with some obvious limitations such as synonyms or multi-word city names (see Section 5.2 for possible workarounds and improvements).

VI. DISCUSSION

In this final section, fulfilled requirements, the evaluation phase with several partners, achieved results, known limitations and possible plans for future work are presented.

A. Evaluation

The GeCCO (German Crafts and Craftsmen Ontology) as well as the tool were evaluated in cooperation with the Chamber of Industry and Commerce of the Stuttgart Region [19] and several regional Chambers of crafts (Handwerkskammer, i.e. organizations that federate craftsmen in a local area), in particular the Handwerkskammer Reutlingen [20], which provided a large test database containing over 10.000 craft service names and contact data.

These prerequisites allowed evaluating the suitability of the tool's design principle, the combination with the ontology and the scalability of the architecture with a dataset of a reasonable size. The evaluation partners were very satisfied with the new search possibilities offered by the tool, and discussions for the realization of productive implementations of the tool are currently in progress.

By using well established graph visualization techniques, it was possible to build a tool allowing non-specialized users to take advantage of a powerful craft services ontology to perform efficient semantic search activities without knowledge in semantic Web technologies nor specific knowledge in craft services.

The semantic search framework presented integrates an ontology navigator with a more classical text-based search, and allows the users to take advantage of both traditional search techniques and of graph navigation, so as to easily find the nearest craftsmen as well as get an overview of the crafts men professional environment.

Since the implementation makes advantage of USDL, the development activities and test results supported the work of the USDL W3C incubator group [16], to help improving and standardizing USDL. The use of USDL also guarantees the possibility to integrate the results in another tool, such as a service marketplace.

TABLE II. FULFILLMENT OF THE REQUIREMENTS

Requirement	Fulfilled	Comment
R1. No need for semantic web knowledge	Yes	Simple keyword-based queries
R2. No need for domain-specific knowledge	Yes	Users do not need to know about craft services since damaged objects are also integrated in the ontology
R3. Navigation across domain classes and relations	Yes	Graphical ontology browser, integration of synonyms in the ontology
R4. High precision of search results	Yes	Exact matching with ontology class
R5. Possibility to integrate in larger framework	Yes	Use of USDL for craft services descriptions

TABLE II summarizes the requirements defined in Section 3 and how they were fulfilled with the search framework we introduced.

B. Limitations

The most obvious limitation of the current implementation lies in the limited test data available, limited to the area of the city of Reutlingen. Although this does not impact the validity of the overall approach, including larger data sets is an important improvement factor.

Another improvement could be moving away from a Flash based framework, so as to allow the usage of the tool on the more and more common platforms which do not support Flash, such as current tablets or smartphones. The Protovis Framework [21] offers similar functionality with full HTML-based technologies and would possibly provide a satisfying replacement for Prefuse Flare.

Finally, concerning the Web search tool and the one-box search input currently used, limitations come from the word-based way used by the software to distinguish places, objects, craft services or craft service company names. For example, if the user enters "Bad Cannstatt Glaser", the tool will tend to match "Bad" to the object "Bath (tub)" and "Glaser" to the service "Glazier", leaving Cannstatt as a potential place or name. The correct matching would have been "Bad Cannstatt" as a place (a suburb of Stuttgart, Germany). To solve this problem, a possible solution would be to iteratively try to match unique words, then two or

 ISBN: 978-1-61208-280-6

three-word combinations, and then choose for each matched item the most probable one. A complete algorithm for matching text-based queries to ontology classes and relations has been introduced in [17]. These approaches would however require, especially for geocoding, a solution offering a very high performance as many slightly different queries would be sent for each user request.

C. Future Work

The interest and benefits of a search tool are highly linked to the quality and completeness of the data that can be found by using it. Thus, one of the main goals of the further development of the search framework is the inclusion of larger datasets, probably by cooperating with additional Chambers of Crafts, so as to achieve a much higher recall over larger regions of Germany. Performance was not an issue in tests with a small user group and a moderate dataset. The search framework presented in this paper can be adapted to scale correctly, using a distributed architecture or simply with more powerful hardware.

Furthermore, including more data in the background could help further develop GeCCO, e.g., with regards to the lists of damaged objects which currently is limited to the most common objects concerned by property claims management.

Reasoners could be used to better search the ontology. Several reasoners were investigated and used to assess the integrity of the GeCCO onotology (as mentioned in [1]), but they could also be used at search time for complex queries.

In order to complement the already-performed evaluation with professional users and a limited number of non-specialists, a larger evaluation and test phase would help guarantee the effectiveness of our approach.

Finally, it can be noted that an advantage of our solution is that it can easily be applied to other domains than craft services, provided that an ontology describing the domain and appropriate data can be used or created.

ACKNOWLEDGMENT

The work published in this article was partially funded by the openXchange project of the German Federal Ministry of Economy and Technology under the promotional reference 01MQ09011.

REFERENCES

[1] A. Horch and M. Kintz, GeCCO: German Crafts and Craftsmen Ontology – A Common Crafts Ontology. In: Proceedings of the 8th International Conference on Web Information Systems and Technologies (WEBIST). Porto, Portugal, pp. 355-360 (2012)

[2] E. Christensen, F. Curbera, G. Meredith, and S. Weerawarana, Web Services Description Language (WSDL) 1.1. (2011) Available online under http://www.w3.org/TR/wsdl (retrieved March 2013)

[3] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou, Ontology visualization methods - A survey. ACM Comput. Surv. 39, 4, Article 10 (October 2007)

[4] OntoGraf, http://protegewiki.stanford.edu/wiki/OntoGraf (retrieved March 2013)

[5] Protégé, http://protege.stanford.edu/ (retrieved March 2013)

[6] NavigOWL, http://protegewiki.stanford.edu/wiki/NavigOWL (retrieved March 2013)

[7] OntoStudio, http://www.ontoprise.de/en/products/ontostudio/ (retrieved March 2013)

[8] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J.J. van Wijk, J.J., J.-D. Fekete, and D.W. Fellner, Visual Analysis of Large Graphs. 12th Joint Eurographics/IEEE-VGTC Symposium on Visualization, pp. 1719-1749 (2010)

[9] J. Cardoso, A. Barros, N. May, and U. Kylau, Towards a Unified Service Description Language for the Internet of Services: Requirements and First Developments. In: Proceedings of the 2010 IEEE International Conference on Services Computing, Miami, Florida, pp. 602-609 (2010)

[10] German Crafts Code (*Handwerksordnung*), http://www.gesetze-im-internet.de/hwo/ (retrieved March 2013)

[11] MyHammer, http://www.myhammer.com (retrieved March 2013)

[12] Project openXchange, http://www.openxchange-projekt.de (retrieved March 2013)

[13] A. Horch, M. Kintz, F. Koetter, T. Renner, M. Weidmann, and C. Ziegler, openXchange: Servicenetzwerk zur effizienten Abwicklung und Optimierung von Regulierungsprozessen bei Sachschaeden. Fraunhofer Verlag, Stuttgart (2012)

[14] Prefuse Flare Framework, http://flare.prefuse.org/ (retrieved March 2013)

[15] M. Kintz and J. Finzen, A Simple Method for Mining and Visualizing Company Relations Based on Web Sources. In: Proceedings of the 7th International Conference on Web Information Systems and Technologies (WEBIST), Noordwijkerhout, pp. 597-602 (2011)

[16] K. Kadner and D. Oberle (Editors), Unified Service Description Language XG Final Report. Available online under http://www.w3.org/2005/Incubator/usdl/XGR-usdl-20111027/ (retrieved March 2013). (2011)

[17] Y. Lei, V. Uren, and E. Motta, SemSearch: A Search Engine for the Semantic Web. Proc. 5th International Conference on Knowledge Engineering and Knowledge Management Managing Knowledge in a World of Networks, Lect. Notes in Comp. Sci., Springer, Podebrady, Czech Republic, pp. 238-245 (2006)

[18] P. Cimiano, ORAKEL: A Natural Language Interface to an F-Logic Knowledge Base. Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems, pp. 401-406 (2004)

[19] IHK Region Stuttgart, http://www.stuttgart.ihk24.de/ (retrieved March 2013)

[20] Handwerkskammer Reutlingen, http://www.hwk-reutlingen.de/ (retrieved March 2013)

[21] Protovis, http://mbostock.github.com/protovis/ (retrieved March 2013)

[22] Google Maps API, https://developers.google.com/maps/ (retrieved March 2013)

[23] Apache Axis2, http://axis.apache.org/axis2/java/core/ (retrieved March 2013)

[24] MySQL, http://www.mysql.com/ (retrieved March 2013)

[25] Apache Tomcat, http://tomcat.apache.org/ (retrieved March 2013)

[26] OWL Web Ontology Language, http://www.w3.org/TR/owl-features/ (retrieved March 2013)

[27] M. Paolucci, T. Kawamura, T.R. Payne, and K.P. Sycara, Semantic Matching of Web Services Capabilities. In: Proceedings of the First International Semantic Web Conference on The Semantic Web (ISWC '02), Ian Horrocks and James A. Hendler (Eds.). Springer-Verlag, London, UK, UK, pp. 333-347 (2002)

 ISBN: 978-1-61208-280-6

A Requirements Model for Composite and Distributed Web Mashups

Vincent Tietz, Oliver Mroß, Andreas Rümpel,
Carsten Radeck and Klaus Meißner
Technische Universität Dresden,
Faculty of Computer Science, Germany
{vincent.tietz,oliver.mross,andreas.ruempel,
carsten.radeck,klaus.meissner}@tu-dresden.de

***Abstract*—Mashups have recently become popular due to visual composition metaphors and loosely coupled widgets that encourage the fast implementation of situational web-based applications. However, the mashup development process is still challenging for end-users, since they are compelled to retrieve and combine adequate components for their requirements manually. Therefore, we propose a novel model-based requirements-driven mashup design and composition method. It benefits from the use of semantics-based domain vocabulary accompanying the whole mashup development process. In this paper, we present an ontology for specifying requirements for web mashups, which is suitable for deployment in multi-user and multi-device scenarios. To this end, we employ a distributed multi-user scenario and outline the proposal's benefits in a model-driven web mashup composition process.**

***Keywords*—Web Engineering, Requirements Modeling, Mashup Engineering, Mashups**

I. Introduction

Presentation-oriented *mashups* have evolved from simple data-driven aggregation of feeds to complex applications composing Web- and UI-based building parts. Compared to other software systems, mashups are rather small applications providing rich user interfaces. Regarding mashup development, simplicity and re-usability are important demands. Therefore, technical details are usually hidden by black-box components providing declarative externalization of their functionality. However, in many of current mashup development environments, the composition is mainly driven by manual selection of components and low-level assembly by connecting events and operations. With the increasing number of Web-based services and components, the development of applications with current mashup platforms becomes a cumbersome task, especially for end-users [1].

Moreover, in modern application scenarios, the context of use is not limited to a single platform or user [2]. People collaborate via their personal devices in a mobile and ad-hoc manner. To achieve abstraction from platforms and service implementations, there is a need for model-based integration concepts, such as provided by CRUISe [3]. These *composite mashups* provide the basis for realizing platform-independent mashup composition as well as distributed and collaborative scenarios that need to be considered in new development methods. Therefore, we argue that requirements-oriented development is needed to cope with such complex scenarios and to abstract from such composition details [4]. Since, a formal requirements specification for mashups is still not available [1], we propose a task-based requirements model as the foundation for stuctured, semi-automatic mashup development and human-centered design of mashup requirements.

The remaining paper is structured as follows. In Section II, we introduce a collaborative travel planning scenario to motivate the need for the representation of distribution and collaboration requirements. Since we propose an extended task metamodel, related work in the field of task-based modeling is discussed in Section III. In Section IV, we present our requirements model for mashups, and its semantic representation. In Section V, we outline the benefits for the mashup development process. Finally, we discuss our results and provide an overview about the implementation of the requirements model as well as transformation process in Section VI and conclude the paper in Section VII.

II. Collaborative Travel Planning Scenario

To illustrate the modeling approach, we introduce a social travel planning scenario, wherein several participants can synchronously vote for a list of travel offers in a distributed co-located multi-device environment. The result is a shared list of rated offers, which is calculated based on the vote from each planning participant. As illustrated in Figure 1, the abstract task tree consists of two phases: the personal offer research (left) and the collaborative rating phase (right). In the first phase, each planning participant creates a personal list of travel offers. This implies the selection of destination locations, e. g., using a geographical map UI component, and one or more offers provided in this region, e. g., using a list component. For example, a participant could choose a bicycle tour in Ireland or an adventure trip to New Zealand.

In the second phase, each participant shares the personal list of offers with all relevant personal devices (smartphone, tablet PC) and shared devices (Smart TV) in a collaborative setting. Next, each device merges the lists in order to visualize the entries and to enable each participant to rate and rank all offer candidates in a synchronous and collaborative manner. To create a common idea of the ranking result, another subtask of the main ranking task – the shared ranking visualization – is executed in the context of a public Smart TV (shared device in the meeting room). After a participant has ranked an entry of the shared list via his personal device, the ranking is synchronized with the shared ranking visualization using the Smart TV.

 ISBN: 978-1-61208-280-6

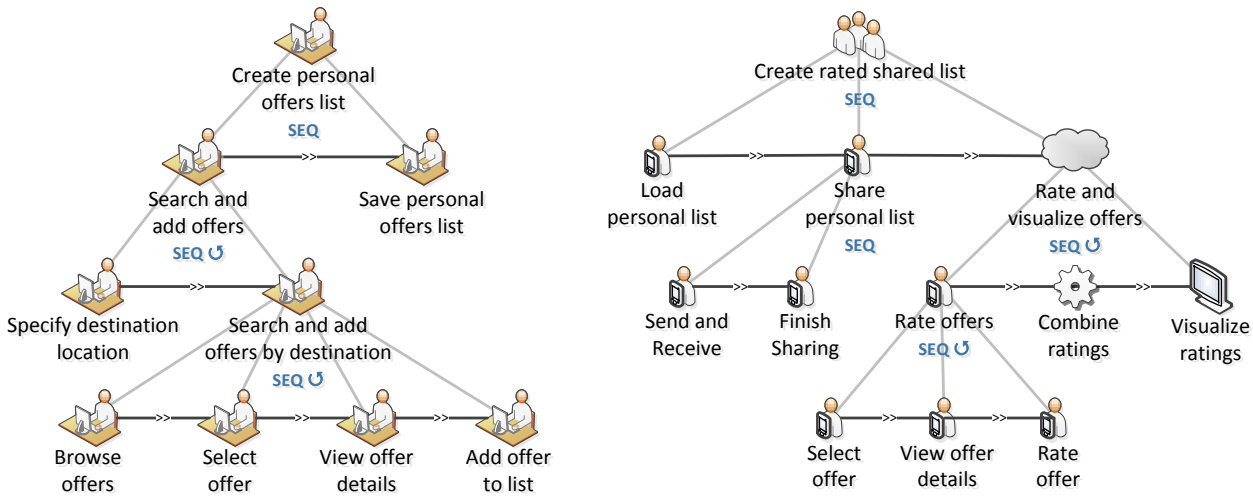


Figure 1. Collaborative multi-device rating scenario

The scenario demonstrates that collaborative tasks are executed in close relation to each other, but in different execution contexts. To create a collaborative application from the task model, additional information and conditions, such as the number of role instances, i. e., collaborators, of the required execution context, is needed. Hence, in the remainder of this paper we present methods to formalize these requirements that we need to determine the binding between a task and its execution context.

III. Related Work

Like in traditional software development, requirements models in the field of Web engineering are, among others, mainly represented by use cases, business process models and task models [5], [6]. Use cases define interaction scenarios between a role and a system in order to achieve a goal, but collaboration and participation among actors cannot be distinguished in UML use case diagrams. Pre- and post-conditions, for example, also need to be added in a textual manner [7]. However, natural language is not suitable to provide a computable requirements model. Further, flow-oriented business process models such as BPMN are well suited for representing workflow requirements. However, both focus on data and system integration and are not suitable to specify individual user requirements [8]. Instead, task models such as ConcurTaskTrees (CTT) [9] are designed for the human-centered specification of user requirements. Because mashup components can be considered as self-contained entities solving user tasks, we argue that task modeling provides both a basis for user-centered requirements modeling and the seamless integration into a model-driven mashup development process.

In the past, several task meta models and notations have emerged with different purposes and degree of formalization and expressiveness. For example, the early Hierarchical Task Analysis (HTA) [10] focuses the hierarchical decomposition of human activities and can be considered as the basis for any subsequent task model. However, tasks are structured by informal plans, goals and actions can be defined on any level of decomposition. GroupWare Task Analysis (GTA) [11] focuses on the collaborative aspect with roles and agents with clear distinction between tasks and actions adopting the activity theory [12]. However, explicit concepts for modeling of distribution are missing. K-MAD [13] is similar to GTA but extends object modeling by supporting abstract and concrete objects. However, their semantics can not be defined. The most prominent task modeling approach is CTT [9] that is used as a starting point in many model-based user interface development (MBUID) approaches. CTT allows the definition of cooperation and communication tasks, which are modeled in collaboration task trees. However, as in all mentioned approaches, distributed and simultaneously performed tasks cannot be expressed explicitly. Finally, there is a lack of semantics and continuous application development, since presentation objects, which are in our case mashup components, need to be selected and integrated manually.

The provision of ontology-based modeling is mainly considered in the field of semantic web services. For example, OWL-S [14] provides an ontology-based specification of web services and templates to initiate automatic search and composition. Similar to that, OWL-T [15] is supposed to support task-based description and discovery. However, the focus is on technical web services with search profiles, grounding facilities, process definitions and result modeling. We adopt some principles such as ontology modeling and semantic matching, but in contrast, we strive for lightweight and user-centered requirements specification focusing on presentation-oriented mashup components.

To generate an application from collaborative task models, several sub models are needed to describe the execution context of the corresponding task. Pribeanu et al. [16] present a hybrid approach to describe context-sensitive and context-insensitive tasks in a single task model. Thereby, CTT is combined with additional model elements such as decision graph nodes and arcs that describe the context of use of each subtask. However, this approach focuses only on how context information can be integrated into the task model, but it does not clarify the context information that is relevant to

 ISBN: 978-1-61208-280-6

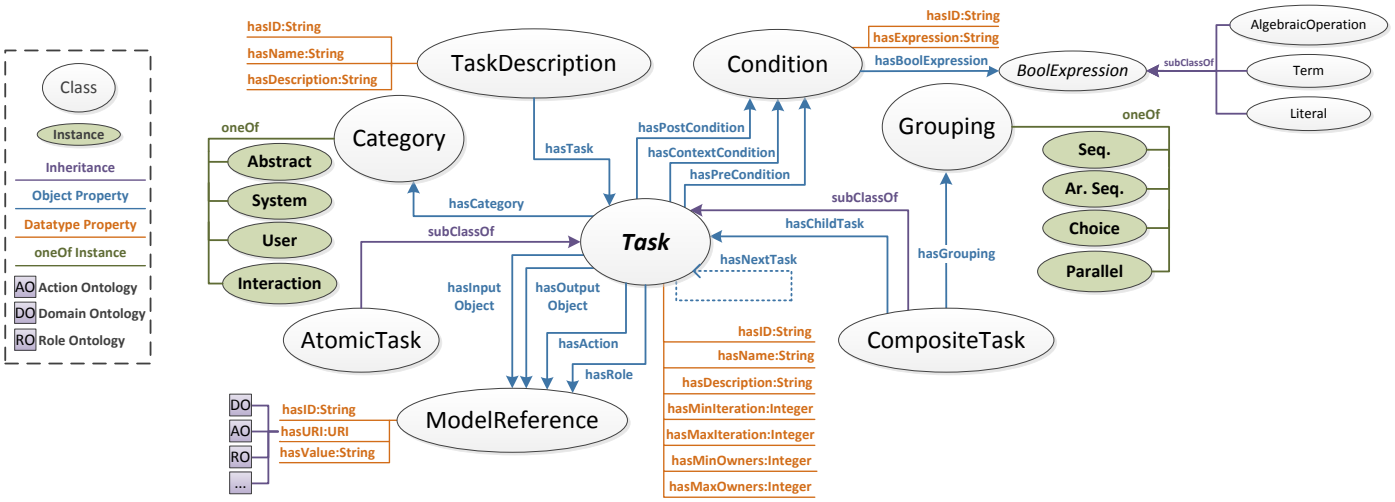


Figure 2. Task ontology

perform context-sensitive tasks in a collaborative fashion. The MBUID approach CODAMOS [17] extends DynaMo-AID by an environment model that classifies each available computing device as an interaction resource with input and output channels. The corresponding environment ontology contains concepts that are referenced by the task model. Finally, so-called modality interaction constraints are used to map tasks to a set of potential interaction resources. However, as in all MBUID approaches relying on the CAMELEON reference framework [18], additional abstract user interface (UI) descriptions are required to transform the task model into a dialog and presentation model. We argue that the mashup paradigm allows us to create web applications with sophisticated UI in a more flexible way. In contrast, we focus on the dynamic mapping of task model entities to black box mashup UI components to reduce the development complexity.

Overall, the lack of degree of formalization regarding data and activity modeling as well as the missing expressiveness regarding the execution context impede the use of traditional requirements modeling approaches in distributed mashups. Hence, we introduce an ontology-based task model in the next section to provide a basis for requirements specification and model-based mappings and transformations.

IV. Requirements Model for Web Mashups

We consider task models as a basis for the specification of mashups because of the intuitive representation of composite user goals and tasks [9], the correlation of tasks and mashup components as tools for solving specific user tasks [19] and the correlation between task models and business processes to represent work structure [20]. Further, we follow [2] by extending task models with context modeling facilities to support collaborative and distributed task design. To show how the mapping from task models to mashup applications works, we refer to Section V. In this section, we describe the main concepts of our task model, which is illustrated in Figure 2 and incorporates such context requirements.

A. Modeling Activities

The structuring of human activities is one major goal of task modeling. In general, this is realized by the decomposition of high level tasks to more fine-grained sub tasks. Although the decomposition is provided by all task models, the understanding about the main concepts such as tasks, actions and operations seems to be rather different [12]. Therefore, we propose to apply the activity theory [21] to structure terms and relations in the task modeling domain. Therein, a human activity hierarchy is proposed, consisting of activity, action and operation. An activity is directed to a motive and is performed within a context of environment, other activities or humans. This corresponds to our *Task* concept.

A task, respectively an activity, consists of actions, describing what needs to be performed in order to reach the goal of the activity. In terms of mashups, actions reflect the required functionality of a component to fulfill a specific task. In our task model, this is represented by *Action* that is associated to each *Task*. In contrast to other task models we provide a classification of actions [22] based on literature from the field human-interaction and visualization such as [23], [24] and [25]. According to the systems theory we classify each action as a subclass of *input*, *transform* and *output* actions. A visual representation of a part of this classification is shown in Figure 3. It is notable, that none of the actions is bound to a concrete interaction type or application. Thus, this classification is intended to be an upper ontology for rather domain specific actions. However, all actions can be potentially performed by humans, by components or by both.

Since mashups support interactive and non-interactive tasks, we use the concept *Category* to denote the interaction type, such as in [9]. *System* tasks are exclusively performed by components. For example, “Combine ratings” is modelled as a system task in our previous scenario in Section II. Whereas, *interaction* indicates that an interaction of humans and a UI is required such as “Specify destination location”. User tasks require no interaction with the system, e. g., fetching a sheet of paper. An abstract task groups heterogeneous subtasks, e. g., “Rate and visualize offers”. Task decomposition is realized by

the concept *Composite Task*, which consists of at least two subtasks *(hasChildTask)*. *Atomic Tasks* represent the leaves of the task tree. To define the workflow, respectively the temporal relations between child tasks, we use the concept *Grouping* consisting of *sequence*, *arbitrary sequence*, *choice* and *parallel*. To define the order of tasks in sequences, the following task is defined by *hasNextTask*.

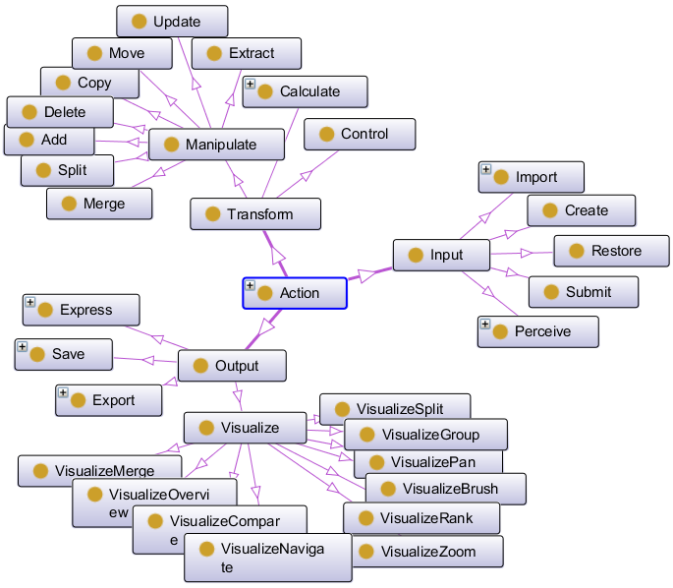


Figure 3. Action ontology

Actions can be very general and usually represent a transformation of a physical or non-physical entity [11]. This is represented by inputs *(hasInputObject)* and outputs *(hasOutputObject)*. The *ModelReference* allows the definition of a data type with the help of a *hasURI*, of a value *(hasValue)* and the instance identification by *hasID*. This also enables the modeling of data flow, wherein output objects of one task can be defined as input objects of another task. In our case, actions, data objects and roles refer to this kind of semantic resources within an action and domain ontology.

Finally, according to the activity theory, operations describe how actions are realized. These reflect the very low-level and modality-dependent interactions of a user such as pressing a key as well as operations of components that need to be performed to provide a needed functionality. However, because we strive for task-based mashup composition and abstraction of implementation details, we do not consider this level of detail in our task model.

B. Modeling Context

To describe the task execution context, we utilize a platform classification that can be used to define context conditions in the task model. It allows the formalization of requirements to perform a task in a specific execution context. We define a platform as a composite of device-, software- and common aspects-specific concepts. To express platform conditions on different abstraction layers we use a vocabulary based on existing classifications such as the *W3C Delivery Context Ontology (DCO)* [26]. Further, we extend the existing modality concept with an additional medium and mode class that are specified as part of the *W3C Extensible MultiModal Annotation Markup Language (EMMA)* [27]. This allows us to characterize the nature of a modality in a detailed fashion, e. g., the description of a *TactileInputModality* with mode *Touch* expresses the necessity of a touch sensitive input device like a multi-touch display or touch-pad.

To allow the description of software-specific requirements in relation to the execution context of a task, the platform model contains corresponding concept elements. The following example will illustrate this. As part of the previous scenario in Section 1, the domain expert defines the atomic task "View offer details" for presenting detail information of a selected travel destination. The UI component should include a video element to visualize the beauty and the richness of the landscape and travel opportunities. To influence the component selection and its distribution (video playing capable device) the domain expert requires a vocabulary to describe the need of the video player capability.

Listing 1 shows an example of a conditional platform statement in SPARQL format [28] to present the expressiveness of our task model. The condition defines constraints for the execution device of a task such as it is capable to play a video. As a consequence, we can use reasoning techniques to improve the quality of the result set during the task-component matching procedure.

```
1  <owl:NamedIndividual rdf:about="
       sample:MobileDeviceCondition">
2      <rdf:type rdf:resource="task:Condition"/>
3      <task:hasExpression rdf:datatype="xs:string">
4          PREFIX p: <http://example.org/platform#>
5          SELECT ?platform
6          WHERE {
7              ?platform p:executedOn ?device.
8              ?platform p:enablesAspect
                   p:BidirectionalCommunication.
9              ?device p:embeds ?display; p:isMobile true.
10             ?display p:supportsModality ?modality.
11             ?modality a p:VisualOutputModality.
12             ?modality p:hasMode p:Image; p:hasMode
                   p:Video.
13             ?display p:hasResolution p:HVGA.
14         }
15     </task:hasExpression>
16 </owl:NamedIndividual>
```

Listing 1. RDF/XML representation of a platform task condition

Besides device- and software-specific concepts, we added a common *aspect* class, e. g., providing a communication classification. In our scenario, the sharing subtask implies a mashup component that should present all traveling planner's ratings in a visual manner (e. g., a list of shared ratings per each travel opportunity). To execute the shared rating list component, the component's platform should provide the capability to receive each favorites list and synchronize with updates from other traveling planner's devices. The collaboration aspect is represented by the role assignment in each task *(hasRole)* and specification how many owners the task might have *(hasMinOnwers and maxOwners)*.

V. Leveraging the Requirements Model

Due to the degree of formalization, the proposed task model is an appropriate basis for further mappings and transformations within a model-driven mashup development process

 ISBN: 978-1-61208-280-6

as already outlined in [4]. For this, we rely on a semantic component model as well as a distributed mashup runtime environment. The main concepts for this purpose stem from the model-based mashup composition approach provided by CRUISe [29], [30]. Therein, mashup components and their interfaces are described by the *Semantic Mashup Description Language (SMCDL)* that enables the semantic annotation of component's meta-data and their interfaces. Figure 4 shows the parts of the SMCDL as well as their relations to functional and data semantics that are defined in domain models. This is the foundation for the universal and semantic description of any web resource such as UI widgets or RESTful web services. Therefore, the SMCDL is the basis for the derivation of an executable mashup application from the requirements defined in our task model.

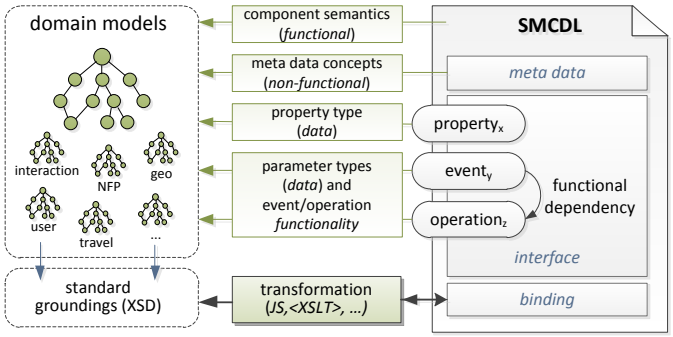


Figure 4. Semantic annotations used in SMCDL [30]

The transition consists of two parts: (1) the discovery of adequate components for each task in the task tree, and (2) the generation of a composition model that can be executed by the CRUISe runtime. The matching algorithm is already described in detail in [19], and is therefore not in the scope of this paper. However, possible mappings for a part of our travel planning scenario are shown in Figure 5. The matching algorithm returns a set of rated mashup components for each leaf of the task tree. In our case, the task "Specify destination location" is mapped to a map component (a), whereas "Browse offers" (b) and "Select offer" (c) to one list component, since it is possible that one component supports more than one task. Further, the task "View offer details" could require additional components to work such as an image database to show further pictures (d). Finally, it is possible that no component can be found for the tasks in the task tree. Unfortunately, this implies either the re-design of the task model or the implementation of new components.

To build a mashup from the component proposals, we adhere to the platform-independent composition model [29]. The composition model consists of five main parts: (1) the conceptual model that contains mashup components to be integrated, (2) the communication model that interconnects components with the help of communication channels, (3) the layout model that describe the arrangement of components on the screen, (4) the state model (previously called screenflow model) that describes the relation of components and screens, and finally (5) the adaptivity model that describes adaptation aspects. The adaptation model is omitted so far, because adaptation aspects affect only composition aspects such as component's reconfiguration, component mediation and layout

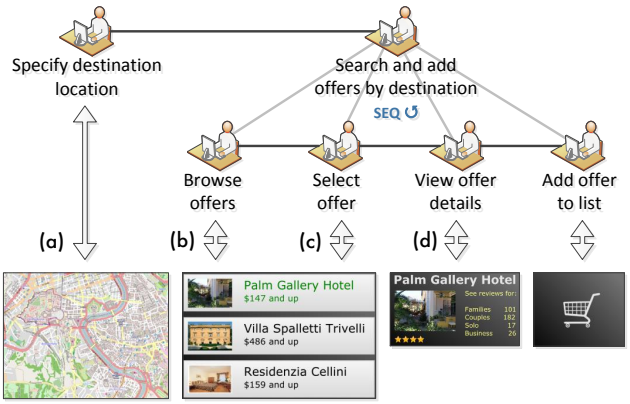


Figure 5. Mapping of tasks and components

adaptation according to the width and height of the screen. However, the task model is of higher abstraction and is not intended to consider such runtime-specific aspects.

Since, the transformation process cannot be explained in detail here, we just provide a brief overview that is illustrated in Figure 6. Therein, a task model is represented by the tasks $T1$ to $T5$ that are connected by model references, such as *Hotels* as output of $T3$ and input of $T4$, and their grouping, e. g., parallel or sequential. Further, conditions could be defined in the task model. The creation of the composition model is then performed as follows.

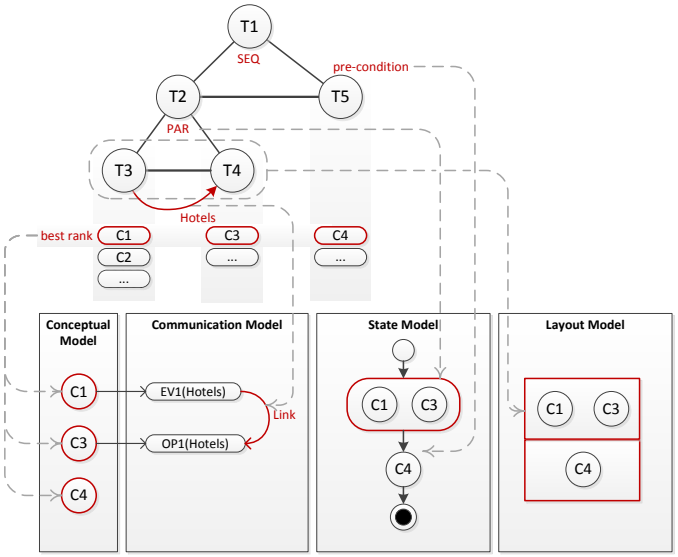


Figure 6. Overview of the transformation process

The conceptual model is created with the help of each component description that is returned by the matching algorithm. Further, the components are reviewed, if they can be connected with each other, i. e., that their semantic descriptions are compatible. This is determined by a matching degree of the parameters of the events and operations. If no matching can be found, it is required to look for further components to make the composition work. For example, additional service components could be needed to fetch some information from external

databases. To fill this gap, the semantics-based discovery of CRUISe and mediation techniques [30] are utilized. When the components are identified, they are collected in the conceptual model to build any of the subsequent model parts. In Figure 6 the selected components are $C1$, $C3$ and $C4$.

Based on the previous results of matching of events and operation, the communication model is derived by both the task model and the component descriptions. First, from the correlation of components to tasks and the definitions of input and output references in the task model, we derive which components need to communicate with each other. Second, the annotation of events and operations allows us the mapping to the interface signatures. For example, in Figure 6, the model reference *Hotel* will result in a channel between component $C1$ and $C3$ because $C1$ is related to the task $T3$ which has a model reference to task $T4$. Finally, the component $C3$ is related to $T4$. Therefore, both components should have a common communication channel.

The state model is derived by the grouping attribute of parent tasks and available conditions. For example, the sequence attribute denotes that subsequent components can only be activated if the current component provides the data according to the definition task so that the task can be considered as finished. However, parallel tasks denote that the components are visible and active all the time. For example, $C1$ and $C3$ work in parallel, but since $T5$ is a subsequent task, the related component $C4$ is only activated if $C1$ and $C3$ finished their computation. In fact this means, that related events provide the required data during the mashup runtime.

Finally, the layout model is derived by the allocation of tasks in task tree and the relation of components regarding their tasks. For example, components that correlate with task neighbors are placed nearby on the screen. However, the algorithm can only provide proposals for the layout and the user is afterwards able to configure this as he or she likes. Further, we introduce a card and a tab layout to hide or disable components according to the grouping definition (e. g., sequence, choice and parallel) and the provided conditions in the task tree.

To consider the platform requirements, we extended the composition model by query-based declarative entities (component distribution description). Each define a mapping between the associated component instance and it's platform condition as shown in Listing 1. The quantity of task owners as well as the role definition *(hasRole)* is also considered. However, if necessary, the definition of capabilities per role needs to be done manually after the generation process, since these information is usually not available in the role ontology. The concrete number of distributed component instances is derived from the quantity of task owners that perform a task simultaneously. Details about their communication behavior are derived from the communication model, but can be configured by the composer too, such as the synchronization interval threshold or their communication direction. It depends on the application scenario and can not be discussed in more detail at this point.

When the composition model has been finished by the composer and validated, the application is finally deployed. The model is directly interpreted by a distributed runtime environment that integrates all referenced components from the repository within the execution context of different client-side runtime environments. Further, it handles and coordinates the component's life cycle, communication and their distributed execution. A part of the resulting application (executed in the "personal offer research" according to our scenario in Section II) is shown in Figure 7.

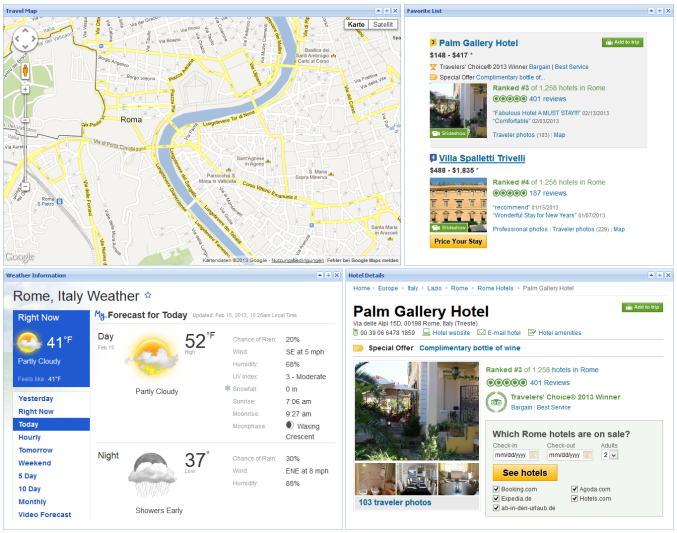


Figure 7. A partial mashup from the scenario

VI. Implementation and Discussion

In this section, we sketch some implementation details of the requirements model and the transformation process, and finally, we discuss our findings.

A. Implementation

1) The Requirements Model: We implemented the requirements ontology in OWL and published an initial version of the proposed meta model on the Web [31]. Listing 2 shows a simple instance of the "Share personal list" task (Figure 1) as an extract in *Terse RDF Triple Language (Turtle)* syntax. The full example is also available in the Web [32]. The first task is an OWL individual of *CompositeTask*, which encompasses two atomic child tasks "Send and recieve" and "Finish sharing" (lines 8-9). The task has an input object that references the semantic concept *favorites list*. Further, the *hasInputObject* relation indicates the count of inbound data channels. Besides the input object, the resulting data of the task is represented by the output object. A context condition is defined in line 10. It refers to the condition that is defined in Listing 1. The collaborative aspect is modeled by the use of the *hasMinOwners* and *hasMaxOwners* properties. The task and its subtasks should be performed in parallel by maximal four and minimal two task owners (travel planners) in the context of their personal mobile devices (lines 16-17).

```
1  :SharePersonalList
2   a activity-tasks:CompositeTask, owl:NamedIndividual ;
3   activity-tasks:hasID "SharePersonalList"^^xsd:string ;
4   activity-tasks:hasName "Share Personal List"^^xsd:string
        ;
5
6   activity-tasks:hasChildTask
```

 ISBN: 978-1-61208-280-6

```
    :FinishSharing ,
    :SendAndReceive ;

 activity-tasks:hasContextCondition
    :sample:MobileDeviceCondition ;

 activity-tasks:hasGrouping
   activity-tasks:Sequence ;

 activity-tasks:hasMaxOwners 4 ;
 activity-tasks:hasMinOwners 2 ;

 activity-tasks:hasInputObject :Favorites ;
 activity-tasks:hasOutputObject :Favorites.
```

Listing 2. Turtle representation of the modeled sharing task

The OWL model is mainly for storing and reasoning purposes. It is obvious that end-users will not create such models without appropriate tool support within the mashup environment. For now, we also created an EMF Ecore [33] representation that allows us to specify a domain specific language (DSL) with the help of EMFText [34]. To provide an example, the same task in Listing 2 is also shown in Listing 3, but using task modeling DSL. The DSL is a more convenient way, and allows the faster creation of text-based task models for testing purposes. Since, the expressiveness required by our task model can be provided by OWL as well as Ecore, the DSL and the OWL representation can be easily transformed into each other.

```
TaskModel TM "Travel planning scenario"

Tasks
    Interaction Task SharePersonalList "Share Personal
         List" {
        action Share
        input Favorites
        output Favorites
        minOwners 2
        maxOwners 4
        Sequence
            Interaction Task FinishSharing "Finish
                 Sharing" {...}
            System Task SendAndReceive "Send and Receive"
                 {...}
    }

References
    Reference Share URI "actions.owl#Share"
    Reference Favorites URI "travel.owl#FavoritesList"
```

Listing 3. The sharing task as DSL

2) The mashup composition process: The instances of the task model are managed by a Java-based *Task Repository (TaRe)* that provides a service interface to store and load task models. It imports RDF/XML as well as DSL-based task models. Further, the TaRe interacts with the CRUISe component repository to find and rate mashup components. The TaRe is also responsible for the generation of mashup composition proposals as described before. Task models are represented internally by an ontology model, facilitating the semantic web framework Jena [35] which allows us to apply SPARQL queries and reasoning techniques.

B. Discussion

As depicted with the previous examples, our task model is applicable to describe simple, more complex and even collaborative scenarios. Besides the travel planning scenario, we also modelled other applications, such as a stock exchange mashup and business trip requests. Further, we evaluated the expressiveness of the task model with the help of the basic workflow patterns [36] and conclude that most of them are supported to model relevant aspects of work. Since we derived the task model from traditional task modeling approaches, we consider that our task model has at least the same expressiveness. Moreover, since the clear semantics provided by the task and action ontology, we are confident of the opportunities for model-driven mashup development. Finally, we showed that a transformation process from a task model requirements specification is feasible.

However, determining the optimal components for single tasks and finding the optimal glue code with respect to interface compatibility and to the task model is of high algorithmic complexity. Although the provided transformation process can produce a draft of the composition model, there are still options and alternatives which might be preferred differently by the other users. Therefore, the user also needs to be able to explore these alternatives and to control the transformation process. Further, composition details such as layout and styles are not covered by the task model explicitly. This is not very surprising as the task model is intended to be an requirements model that abstracts from composition details. Therefore, the user should be able to refine the resulting model manually. To enhance the generation process, we plan to store compositions resulting from a task model and to let users rate them. Based on this, we will be able to generate more detailed composition proposals. Finally, our experience is that the component matching and the composition process is time consuming, depending on the amount of tasks and components. Therefore, we propose to save matching results for later reuse to save calculation time.

VII. Conclusion and Future Work

In this paper, we proposed an ontology-based task metamodel for composite mashups with special aptitude to requirements specification for distributed and collaborative scenarios. To this end, a mashup composer can structure the intended behavior of the application by decomposing corresponding tasks, while simultaneously defining conditions on the distribution, platform or devices. In contrast to other existing approaches, our task models are able to provide clear semantics regarding the actions to be performed within tasks, data objects and conditions. We illustrated the applicability of the presented metamodel by instantiating it for a social decision support scenario.

Future work includes the evaluation of the transition process from a task model to the mashup composition model with the help of a user study. In order to do this, we are currently working on a Web-based task model editor to provide an abstract view on ontology-based task modeling, making it feasible for developers without RDF/XML knowledge or advanced programming skills. Since task modeling activities depend on the application domain, constraints regarding the execution of mashups and their components such as time-out conditions or other quality requirements may be also specified along the task definitions. Conclusively, we claim that task-based mashup development will bring the mashup application building paradigm to a broader audience. Task-based mashup development implies new opportunities for the usage of task

 ISBN: 978-1-61208-280-6

knowledge during the composition and runtime, e. g., in order to optimize tasks and dynamic component integration.

ACKNOWLEDGEMENT

The work of Vincent Tietz is granted by the European Social Fund (ESF), Free State of Saxony and Saxonia Systems AG (Germany, Dresden) within the project eScience, contract no. ESF-080951807. The work of Oliver Mroß is granted by the ESF, Free State of Saxony, contract no. ESF-080951831. Same applies for Carsten Radeck who is funded under the contract no. ESF-080951805.

REFERENCES

[1] V. Tietz, A. Rümpel, C. Liebing, and K. Meißner, "Towards requirements engineering for mashups: State of the art and research challenges," in *Proceedings of the 7th International Conference on Internet and Web Applications and Services (ICIW2012)*, Stuttgart, Germany, 2012.

[2] J. Vanderdonckt, "Distributed user interfaces: how to distribute user interface elements across users, platforms, and environments," *Proc. of XIth Congreso Internacional de Interacción Persona-Ordenador Interacción*, pp. 3–14, 2010.

[3] 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 & Service (iiWAS'10)*, 2010, pp. 413–420.

[4] 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*, 2011.

[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] 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.

[7] 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.

[8] K. Sousa, "Model-driven approach for user interface: business alignment," in *EICS '09: Proceedings of the 1st ACM SIGCHI symposium on Engineering interactive computing systems*. New York, NY, USA: ACM, 2009, pp. 325–328.

[9] F. Paternò, C. Mancini, and S. Meniconi, "ConcurTaskTrees: A diagrammatic notation for specifying task models." Chapman & Hall, 1997, pp. 362–369.

[10] J. Annett and K. Duncan, "Task analysis and training design," Hull Univ. (England). Dept. of Psychology., 1967.

[11] M. van Welie, G. C. van der Veer, and A. Eliëns, "An ontology for task world models," in *5th Int. Worksh. on Design, Specification, and Verification of Interactive Systems (DSV-IS)*, 1998.

[12] Q. Limbourg and J. Vanderdonckt, "Comparing task models for user interface design," in *The handbook of task analysis for human-computer interaction*. Lawrence Erlbaum Associates, 2003, pp. 135–154.

[13] S. Caffiau, D. L. Scapin, P. Girard, M. Baron, and F. Jambon, "Increasing the expressive power of task analysis: Systematic comparison and empirical assessment of tool-supported task models," *Interacting with Computers*, vol. 22, no. 6, pp. 569–593, 2010.

[14] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. N. M. Paolucci, B. Parsia, T. Payne, E. Sirin, N. Srinivasan, and K. Sycara. (2004, November) Owl-s: Semantic markup for web services. W3C. [Accessed: 2013-04-16]. [Online]. Available: http://www.w3.org/Submission/OWL-S/

[15] V. X. Tran and H. Tsuji, "Owl-t: An ontology-based task template language for modeling business processes," in *Software Engineering Research, Management Applications, 2007. SERA 2007. 5th ACIS International Conference on*, 2007, pp. 101 –108.

[16] C. Pribeanu, Q. Limbourg, and J. Vanderdonckt, "Task modelling for context-sensitive user interfaces." in *DSV-IS*, ser. Lecture Notes in Computer Science, C. Johnson, Ed., vol. 2220. Springer, 2001, pp. 49–68. [Online]. Available: http://dblp.uni-trier.de/db/conf/dsvis/dsvis2001.html#PribeanuLV01

[17] T. Clerckx, C. Vandervelpen, and K. Coninx, "Task-based design and runtime support for multimodal user interface distribution," in *Engineering Interactive Systems*, ser. LNCS. Springer Berlin / Heidelberg, 2008, vol. 4940, pp. 89–105.

[18] L. Balme, R. Demeure, N. Barralon, J. Coutaz, G. Calvary, and U. J. Fourier, "CAMELEON-RT: A software architecture reference model for distributed, migratable, and plastic user interfaces," in *EUSAI*. Springer-Verlag, 2004, pp. 291–302.

[19] 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.

[20] H. Trætteberg, "Modeling work: Workflow and task modeling," in *CADUI*, 1999.

[21] V. Kaptelinin, *Context and Consciousness: Activity Theory and Human-Computer Interaction*. The MIT Press, 1996, ch. Activity Theory: Implications for Human-Computer Interaction, pp. 103–116.

[22] V. Tietz. (2011, June) Actions Ontology. Faculty of Computer Science, Technische Universität Dresden. [Accessed: 2013-04-16]. [Online]. Available: http://mmt.inf.tu-dresden.de/models/1.11/actions.owl

[23] R. Springmeyer, M. Blattner, and N. Max, "A characterization of the scientific data analysis process," in *Proceedings of the 3rd conference on Visualization'92*. IEEE Computer Society Press, 1992, pp. 235–242.

[24] D. Gotz and M. X. Zhou, "Characterizing users' visual analytic activity for insight provenance," *Information Visualization*, vol. 8, pp. 42–55, 2009.

[25] G. Mori, F. Paternò, and C. Santoro, "CTTE: Support for developing and analyzing task models for interactive system design," *IEEE Trans. Software Eng.*, vol. 28, no. 8, 2002.

[26] R. Lewis and J. M. C. Fonseca, "Delivery context ontology," World Wide Web Consortium, Working Draft WD-dcontology-20080415, April 2008.

[27] M. Johnston, P. Baggia, D. C. Burnett, J. Carter, D. A. Dahl, G. McCobb, and D. Raggett, "EMMA: Extensible MultiModal Annotation Markup Language," W3C Recommendation, 2009. [Online]. Available: http://www.w3.org/TR/emma/

[28] E. Prud'hommeaux and A. Seaborne, "SPARQL query language for rdf," *W3C Recommendation*, vol. 4, pp. 1–106, 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/

[29] S. Pietschmann, "A model-driven development process and runtime platform for adaptive composite web applications," *Intl. Journal On Advances in Internet Technology (IntTech)*, vol. 4, no. 1, pp. 277–288, 2010.

[30] 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 (Mashups 2011)*. ACM, 2011.

[31] V. Tietz and O. Mroß. (2012, June) Requirements Ontology. Faculty of Computer Science, Technische Universität Dresden. [Accessed: 2013-04-16]. [Online]. Available: http://mmt.inf.tu-dresden.de/models/1.11/requirements-model.owl

[32] O. Mroß. (2012, June) Sample Travel Planning Scenario Ontology. Faculty of Computer Science, Technische Universität Dresden. [Accessed: 2013-04-16]. [Online]. Available: http://mmt.inf.tu-dresden.de/models/1.11/distribution-scenario.owl

[33] D. Steinberg, F. Budinsky, M. Paternostro, and E. Merks, *EMF: Eclipse Modeling Framework 2.0*, 2nd ed., E. Gamma, L. Nackman, and J. Wiegand, Eds. Addison-Wesley Professional, 2009.

[34] EMFText: Concrete Syntax Mapper. DevBoost. [Accessed: 2013-04-16]. [Online]. Available: http://www.emftext.org/index.php/EMFText

[35] Apache Jena. The Apache Software Foundation. [Accessed: 2013-04-16]. [Online]. Available: http://jena.apache.org/

[36] N. Russell, A. H. M. T. Hofstede, and N. Mulyar, "Workflow controlflow patterns: A revised view," BPMcenter.org, Tech. Rep., 2006, bPM Center Report BPM-06-22.

 ISBN: 978-1-61208-280-6

An Approach to Dynamic Discovery of Context-Sensitive Web Services

Victor G. da Silva, Carlos E. Cirilo, Antonio F. do Prado,
Wanderley L. de Souza
Computer Science Department
Federal University of São Carlos (UFSCar)
São Carlos, Brazil
Email:{victor.silva, carlos_cirilo, prado, desouza}@dc.ufscar.br

Vinícius Pereira
Institute of Mathematical Sciences and Computing
University of São Paulo (USP)
São Paulo, Brazil
Email: vpereira@icmc.usp.br

***Abstract*— With the Internet becoming increasingly present in people's lives and the growing availability of Web Services (WS), new challenges have emerged for Software Engineering regarding application development based on composition of highly reusable services within the so-called Service-Oriented Architecture (SOA). One such challenge refers to how to automatically accomplish WS discovery at runtime in order to compose personalized application features that meet particular user's requirements as his/her context of interaction changes. To tackle this issue, we propose in this paper an integrated approach that addresses dynamic WS discovery by combining the traditional WS technology stack with conceptions of Semantic Web on top of the UbiCon, a framework that supports context-sensitive behavior and supplies application with contextual information at runtime. The proposed approach counts on two main elements: (i) an algorithm that performs the dynamic discovery of context-sensitive WS; and (ii) a reusable software architecture that provides a skeleton which enables applying (i). The intention of this approach is to reduce development efforts and increase productivity as it encourages reusing prefabricated WS, as well as to improve software quality whereas applications are assembled as a set of services extensively tested and validated.**

Keywords-Web Services Discovery; Semantic UDDI; Context-Sensitivity; SOAP; RESTful

I. INTRODUCTION

The Web has become one of the biggest media in the world, transforming the way in which people communicate and share content. Since its beginning, the Web has gone through some phases to fit it to recent technological possibilities and to attend new users' requirements. One can mention, for instance, the evolution from the traditional Web to Web 2.0 that is marked by greater interactivity, collaboration and communication among users [1]. This reality has driven to the need for applications that support large amounts of information to different users and devices across multiple contexts of use. This issue has attracted the interest of many researchers over the years. The academic community has proposed different solutions. Some of them, for instance, address frameworks (e.g., [2], [3], [4]) that support development of context-sensitive applications. Meanwhile, the software industry attempts to solve the problem of wide variety of access devices putting into practice the Service Oriented Architecture (SOA) [5] by means of the Web Services (WS) technology, so as to enable integration of heterogeneous systems, fostering greater interoperability.

SOA is widely used in the industry and has the WS as the main way for its realization. The model introduced by the WS technology has a well-defined and structured architecture. In spite of that, some issues still remain for SOA-based software development: One of the main challenges is to perform WS discovery to properly satisfy users' needs, especially at runtime when the context of interaction constantly changes. Semantic Web [6] arises to help out filling this gap, enabling integration among WS and Semantic languages. The Semantic Web is an extension of the Web, which allows computers and humans to work in cooperation, resulting in a better user experience.

It is noticeable that many systems technologies are integrated to provide users with better interactions. On this basis, we present in this paper an integrated approach to enable semantic annotation of WS for context-sensitive Web Services discovery. The approach combines the traditional WS-* stack [7] with Semantic Web technologies to provide developers with an architecture which enables application performing, at runtime, the dynamic discovery of WS in order to foster adaptable behavior based on the users' context of interaction. The framework UbiCon [4] has been applied on our proposal to furnish the contextual information needed to reach context-sensitivity.

The technologies used in the research are presented in the following sections. Section II presents the main concepts regarding the WS technology. Section III presents description languages of Web Services. Section IV broaches concerns about Universal Description, Discovery and Integration (UDDI) and WS discovery. Section V presents conceptions regarding Context and Context-Sensitive Systems. Section VI details our proposal. Finally, Section VIII presents final remarks and further work.

II. WEB SERVICES

WS have emerged as a technology that enables the realization of the Service Oriented Architecture (SOA), whose main purpose involves integrating heterogeneous systems

 ISBN: 978-1-61208-280-6

and delivering applications and features as loosely-coupled services by means of open standards of the Internet (e.g., HyperText Markup Language – HTTP; eXtensible Markup Language – XML; etc). The World Wide Web Consortium (W3C) [8] defines a Web Service as a software designed to support interaction machine-to-machine over a network. WS are independent of the implementation language, which favors interoperability among disparate systems. WS use standard protocols for message exchange. They are identified by a Uniform Resource Identifier (URI), available in a standardized service descriptor, along with the signatures of its features and functions.

There is a well-known architectural model for WS that is based on the interaction of three main components: *Registry*, *Provider* and *Client*. Registry is a repository of WS in which service Providers can post descriptions of their services; Clients, in turn, can query the Registry in order to find out the descriptions of the services they are looking for. The Providers are the ones who furnish the services, which are materialized as Web Services and achievable by some Provider software agent. Clients perform requests to the WS, which involves getting a description of a service from the Registry and then invoking the Web Service functionalities by means of a client software agent for messaging and communication with the Provider agent.

The interaction between these components occurs through some operations: *Publish*, *Find* and *Bind*. When the Provider offers WS, providing a description that is published in the Registry, a *Publish* operation is performed. *Find* operation occurs when a Registry is used by the Client to discover and to retrieve information about services of interest stored in published service descriptors. Finally, the *Bind* operation happens when the Client uses the service description to bind with the Provider and interact with the implementation of the Web Service. WS can be based on SOAP or REST, as explained in the following subsections.

A. Web Services SOAP

The Simple Object Access Protocol (SOAP) [9] is a set of conventions defined by the W3C for exchanging messages that are transmitted and negotiated over a network on the top of the HTTP protocol. Information exchange between the client and the service provider is based on XML format and happens over decentralized and distributed environments in Remote Call Procedures (RPC) style. The SOAP packages the messages to be exchanged in a standard envelope structure, presenting itself as a simplified and lightweight mechanism for exchanging structured information.

By using SOAP, both the application server and the client must be able to interpret the messages structure. This requires the developer to implement appropriate software agents that communicate and understand the protocol's details, resulting in greater efforts to evolve the system. For this reason, using SOAP as a standard technology for WS development has become a deprecated practice. On the other hand, even with recent technology achievements on the WS development, SOAP still has its usage and contribution in software industry and shall be considered in service discovery.

B. Web Services RESTful

RESTful WS are based on the Representational State Transfer (REST) architecture, defined by Roy Fielding [10]. This kind of Web Service has similar characteristics to the SOAP WS, but RESTful ones are lighter and easier to access [11], [12]. A RESTFul Web Service focuses on the service's resources, rather than on its functionalities, and are naturally transferred by the HTTP through its methods GET, PUT, POST, and DELETE. RESTFul WS are gaining momentum, both from the research community and companies [13], which are adopting REST because of its easiness and simplicity of publication, invocation and maintenance [14], [15], [16]. Deploying RESTful WS is quite similar to deploying a dynamic website [11].

Despite of its widespread architectural model, the traditional Client-Provider-Registry architecture for WS is falling into disuse. Client businesses are making off-line agreements and no longer need to query the service *Registry*, because they know in advance the WS' location and how to access them. There are, still, on-line documentation, as well as websites of developers that publish how to access their services, which avoids the use of descriptors and mitigates the need of conventional service registries – which does not adhere to the traditional WS architectural model [11], [12], [17]. This new "model" hinders the dynamic discovery of WS. Since this article deals with service discovery, both SOAP as RESTful must be able to be understood and processed computationally, enabling dynamic discovery. For this end, semantic annotations are used, allowing contextual processing of WS.

III. Description Languages Syntactic and Semantic

The Web Services architectural model provides a means of communication between their organizations, exchanging information through messages and descriptors Web Services. The descriptors describe, syntactically, how to access the service and what the service provides. The W3C has standardized the format descriptor, naming the Web Service Description Language (WSDL) to describe web services in a structured way. According to the W3C [18] WSDL defines an XML grammar and a model for describing network services as collections of communication endpoints capable of exchanging messages. A description of Web Service is a document by which the provider communicates the specifications for the client to invoke the Web Service, thereby defining how the interaction between them should

 ISBN: 978-1-61208-280-6

be, how and where to access, and what the input data and output are.

Still thinking of the architectural model of Web services, service discovery happens through a search for keywords in the *Registry*, preventing the discovery of content. Faced with this problem, the proposal is to use semantics to represent the content of Web Services, enabling the dynamic discovery of services.

The following are the subsections that address the technologies used for syntactic and semantic description.

A. WADL and WSDL Descriptors

The first version of WSDL emerged in mid-2001 and allowed to only describe SOAP-based Web Services. Soon after came the REST-based Web services, RESTful Web Services appointed, using much of the characteristics of the HTTP protocol for messaging between *Provider* and *Client*, which provided an updated version of WSDL to also support the description of Web RESTful Services, called WSDL 2.0.

In addition to the WSDL, which is the descriptor of Web Services standard, there is the WADL proposed by Hadley, to provide a description that is can be processed by machine, and RESTFul resources that is based on HTTP [13]. A large number of web-based companies (Google, Yahoo, Amazon and others) have used RESTful [13] to provide access to its internal data, but use documentation with web sites, instead of using the WADL. WADL is different because it provides an interface in XML, describing an application and not a service, mapping the concepts that form the RESTful paradigm [1].

Even though WADL describes easier and more efficient RESTful Web Services, WSDL describes both SOAP-based Web Services as RESTful, which facilitates the discovery and invocation of both service types, justifying the use of standard WSDL. Finally, it is worth mentioning that a drawback in dynamic discovery of Web Services is that only descriptors include syntactic description of their services, thus needing a semantic description.

B. Ontologies and OWL-S

Ontologies have a very important role in automated discovery and composition of Web Services. They are the ones that describe and give knowledge to the computer that can process and understand the content of Web Services, thus being able to discover, compose and invoke the available services automatically. The most used definition of ontology in the literature on services semantic web is: Ontology is a collection of Web Services that share the same domain of interest and describe how Web Services can be described and accessed [19].

The semantic description is essential for efficient discovery of Web Services, and occurs through ontologies, which are built from a markup language. Research has shown that the ontology OWL-S [20], which extended the Web Ontology Language (OWL) adding semantics, is more efficient in the context of semantic description to Web Services. OWL-S is an ontology for Web Services that lets you discover, compose, monitor and invoke services with self-degree of automation. It is composed of 3 elements: *Service Profile* - announces and discovers services, *Service Model* - describes service operations, *Service Grounding* - provides details of communication protocols and message format.

Joining Web Services with semantic language have Semantic Web Services. The semantic services deal the limitations in current Web Services through the improvement of the description of services, defining a semantic layer, in order to achieve automatic processes of discovery, composition, execution and monitoring [21]. To solve this problem scientists have used four attributes described by the semantic class Service Profile of OWL-S: *Inputs*, *Outputs*, *Preconditions* and *Effect* (IOPE) [22], [23], [24], [25]. Certainly there are functions that are unique to OWL-S, and the main one is the mapping between OWL-S specifications and syntactic specifications of WSDL, which allows the integration of ontologies with Web Services. This mapping is one descriptor to one ontology. Furthermore, the *Service Profile* settings matter are defined in other Semantic Service Profiles and other ontologies, thereby facilitating reuse and avoiding ambiguities [22].

Finally, we have Semantic Web Services that can be searched by content and not just keywords, increasing the power of discovery Web Services in Registry.

IV. UDDI

Web Services should stay available in order to be accessed by any client application present on the Web, using a static or dynamic way. The static mode to access a Web Service is characterized when the service address is provided to the client application to locate and access the service, i.e., the client application needs to know the Web Service address to access and invoke it. The service can also be discovered, characterizing the dynamic mode of discovery a Web Service. Universal Description Discovery and Integration (UDDI) is a platform-independent structure built for dynamically describe and find published services, which are contained in the architectural model of Web Services. According to OASIS [26], UDDI specifications, define a *Service Registry* for Web Services. An UDDI Registry manages information related to service providers, service implementations and their data. *Providers* can use UDDI to advertise their services and *Clients* can use it to find services, which meet their requirements. In this way, UDDI refers to a repository of Web Services provided by companies and their descriptors.

Dustdar et al. [19] asserts that although UDDI Registry and other UDDI based-models have been implemented, they are not widely used, and for the service dynamic discovery

 ISBN: 978-1-61208-280-6

they do not meet the requirements yet. The UDDI-based discovery is performed using keywords. This search includes Web Services with low relevance and ambiguity. To solve this problem there are researches related to dynamic discovery through semantics based on ontologies. The discovery provides semantic search through content and not only by keywords, facilitating the understanding by the computer [20]. The web service discovery depends on the service descriptor, in this case the WSDL. WSDL complements the UDDI standard by providing an uniform way for describing the interface and the connection between *Provider* and *Client*. Since the UDDI specifications do not define a semantic processing for Web Services, UDDI should be extended in order to understand the service semantic context, allowing the dynamic discovery of semantic services within the UDDI.

V. Sensitivity to Context

With the explicit presence of the Internet in daily life, the use of applications that contain a large amount of information is growing rapidly, as well as the need of users to do complex tasks and to process information in a short time, thus creating a challenge to computational systems. The challenge aims to reduce the need of explicit interaction of the user with the system to get what it want [27]. Faced with this problem, researchers have created new systems approach, which aims to understand the context in which the user is in order to assist in the needed actions. These systems are called Context-Sensitive Systems [27] or Context Aware Systems [2]. This article refers to the term Context-Sensitive Systems because it reflects better the semantic of a system which detects context changes, adapts itself and reacts to these changes [3].

A. Context

The context is the key to filter the available information and turn them into relevant information. In the literature there are several definitions of context. According to Vieira [27], Context is everything that involves a situation at a given moment, allowing in the identification of what is and is not relevant to interpret and understand the situation. Santos [3] also makes a distinction between: context and contextual elements, arguing that Contextual Element (CE) is any data, information or knowledge that allows to characterize an entity into domain. As defined in [2], context is any information that can be used to characterize the situation of an entity (e.g., person, place, object, user application). Context involves information that refer to various aspects associated with the operation of the application, such as user, device access, environment, network, among others [28].

Using these ideas, in computer systems Context is an instrument to support communication between systems and their users. From understanding the context the system can, in many circumstances, change its sequence of actions, style of interactions and type of information provided to users in order to adapt to the current needs. Systems using the context to guide actions and behaviors are called Context-Sensitive Systems or Context-Aware Systems [3].

B. Context-Sensitive Systems

Context-Sensitive Systems (SSC) or Context-Aware Systems are computer systems that use context to provide more relevant services or information supporting user tasks, where these tasks are context dependent [2], [29]. There is a big difference between traditional applications and context-sensitive applications. Traditional applications act considering only requests and information provided explicitly by users. Regard to the context-sensitive applications, they mainly differ in the amount of input and output data. The input data can be the explicit information provided by the user, the information stored in knowledge bases contextual, inferred through reasoning, and also those perceived from environment monitoring [3].

To better illustrate the functioning of context-sensitive applications, it is considered, for instance, the case of an application to help medical appointments, which takes into account the patient preferences and context. The patient registers its data with preferences in a database shared by hospitals and clinics. When the patient searches for a doctor to schedule an appointment, the application acts helping him in his actions. The patient can inform their illness or medical specialty required to filter the search for clinics and centers. The system is located next to the appropriate clinical patient, considering its location and preferences, and search for doctors with the specialties required, considering the doctor's schedule to make an appointment. The system sends alert messages for both the patient and the specialist if the query is scheduled successfully, or in the occurrence of a fault.

C. UbiCon

There are many challenges to develop a context-sensitive system. Researches have been developed for the purpose of generating specific tools and methods to support the treatment of these challenges, such as toolkits [2] and frameworks [3], [4]. These approaches aim to meet basic functionality for the management of contextual information so that applications can make use of services, simplifying the development of such systems. The basic features are listed by Vieira [3], setting them into 3 categories:

- Specification Context: handles the context of requirements gathering and modeling the contextual information needed for the application.
- Management Context: indicates how the context will be handled by the system, in terms of the following tasks: acquisition, storage, processing and dissemination of contextual elements.

 ISBN: 978-1-61208-280-6

- Use Context: defines how context influences the behavior of the system and how it will be used effectively.

Considering that any information can be considered context, the information obtained should be relevant. Santos [3] defines a framework called Contextual Elements Management Through Incremental Modeling and Knowledge Acquisition (CEManTIKA) to capture, manage and disseminate computational context, which is used as the base architecture for context extraction in computational work. Faced with the challenges and motivations for building a context-sensitive application, we developed a framework in the same context of research, called Ubiquitous Context Framework (UbiCon) [4]. The UbiCon encapsulates the manipulation tasks context and provides functionalities divided into 4 modules: acquisition, processing, dissemination and adaptation of content, based on context architecture proposed by Santos [3]. Thus arises a motivation to continue this line of research, reusing the results.

VI. Proposal

This proposal can be best viewed considering his major contributions to computing. They are divided into steps, explained below.

A. Architecture

The proposed architecture is shown in Figure 1. This architectural approach proposes the reuse and extension of the architectural model of Web Services, adding knowledge and context sensitivity, obtained by UbiCon framework, and Web Services search for content using ontologies that are constructed from the interpretation and validation of descriptors service. Thus, the architectural model of Web services is extended for the context, compared with the contents of Web services described using semantic, and they can be interpreted automatically by the computer, achieving the object of discovering dynamic context sensitive.

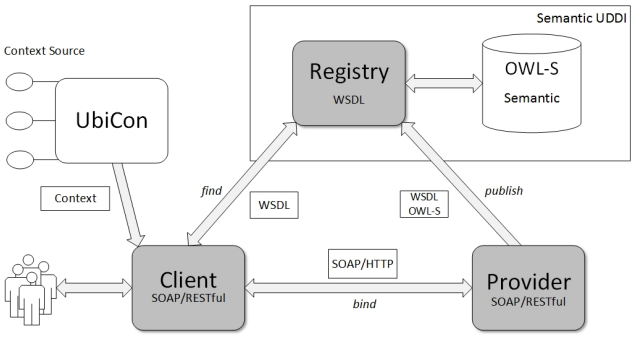


Figure 1. Architectural Model for Dynamic Discovery of Context-Sensitive Web Services.

B. Extending the framework UbiCon

Considering that the service discovery task is dynamic and it varies according to various relevant factors to obtain the contextual information required to be performed, the service discovery must be an adequate automated task. In this way, the UbiCon framework will be extended to retrieve contextual information that will guide the discovery process of services, getting a higher volume of contextual information about the entities associated with the operations of the discovery process of services. Therefore the discovery mechanism may consume contextual information obtained by the modules Acquisition, Processing and Dissemination UbiCon, and perform the discovery of services which meet the contextual variations observed. To allow the automation in discovery tasks, the contextual information involved in the execution of the application (e.g., user profiles, device profiles, network access and SLA) should be available to facilitate reasoning and inference about the same runtime.

One of the biggest challenges in Ubiquitous Computing, cited by Gimson et al. [30], is a description of the delivery context, which is defined as a set of attributes that characterize the capabilities of the access device, the user preferences and other aspects related to the delivery of Web content. To address these problems and in order to achieve the proposed goal, this paper proposes the use of ontologies for describing the characteristics of the delivery context, focusing on the discovery of services. Ontologies allows to express concepts and relationships of domain and turn them computable. This work will build ontologies to represent knowledge about the different profiles and services.

The specification of the ontologies that characterize the Profiles and Services have bases on previous work in building the Extend Internet Content Adaptation Framework (EICAF) [22]. The EICAF is a content adaptation framework for Web applications based on Web services and ontologies. In EICAF, the contextual information about the domain entities considered (e.g., device, user, SLA, network, content) are available through profiles specified in OWL, being easily reused in OWL-S. These ontologies will be refined and extended in this work to meet the requirements of dynamic discovery of services and the context of the application.

C. Discovery of Context-Sensitive Web Services

The Discovery of Context-Sensitive Web Services is the main purpose of the proposed work. It happens when the client wishes to invoke some kind of service that meets their situation or context in which it appears. The *Listener Client*, which is waiting and watching if there have been changes in the context, *searches* the service through the *Semantic Registry*, passing as a parameter the *Context* obtained through of the framework UbiCon. The search occurs through the *Service Profile* class, which extends the *Service* class of ontology OWL-S, that has a high degree of specification [22]. In the *Service Profile* there are features

 ISBN: 978-1-61208-280-6

(in the form of representation of functions that provides the service) that allows the service discovery, which are: *Inputs*, *Outputs*, *Preconditions* and *Effects* (IOPE). Through these representations, the context information and services are compared, returning the descriptions of services that corresponds the sent context, realizing the discovery of services dynamically.

D. Method and Evaluation

The work is being carried out iteratively and incrementally applying an evolutionary process model [31] for the implementation of the activities identified in the previous subsections. In this sense, it follows that for each iteration of the activities, an increase of work is produced. Each increment is made up of parts of the artifacts resulting from each activity iterated until all activities are carried out completely. The first activity was developed and iterated according to the architecture proposed in this paper. In the following iterations, new artifacts are created, tested, refined and evolved until a full version of the project, in order to validate the proposal.

The validation and monitoring will be done using case studies and experiments developed at the end of each interaction, advocating the hypothesis of the work aimed at discovering dynamic context-sensitive services.

VII. Related Works

Tegegne et al. [32] present an architecture/framework for health systems in countries with poor infrastructure. The article covers health systems that manage patient registration, claiming that these systems are the biggest problem in the health area. The article presents a solution, still immature, with SOA, SOC and Web Services technologies to solve these problems. The authors do not specify what type of technology it is used, but they present an approach very similar to the one proposed in this paper. The part of the context that manages patient information and services, makes a junction between these informations, thus, it is discovering services. Presents a specific service repository that can not be reused.

John et al. [33] define in his work a standard RDF with Microformats aiding semantics for RESTful WS. It describes services using annotations in HTML and work in specific cases, but in generic cases it does not work.

Ferreira Filho et al. [34] create an ontology to aggregate with RESTful WS. Their approach uses an extension of OWL-S Grounding and WADL descriptors. The work does not address SOAP services and does not use the standard WSDL.

In this work, RESTFul and SOAP services are discovered based on the descriptions WSDL2 and OWL-S. The UbiCon framework is used for context management. The repository present in this article can be inserted legacy services, this enables the reuse.

VIII. Final Remarks and Future Works

This paper proposes a new approach to discover Web services, using technologies that enable the use of context and semantics which support searching by content and context-sensitive. This proposal allows the reuse of other sources of finding web services, taking as an input the descriptions of the services, which will be converted to OWL-S, adding semantics and allowing the search for content within the registry. The article discusses problems and challenges of the area, approaching a solution and contributing to the state of the art.

Future work aims to improve the semantics of Web services with the addition of WordNet [35], improving the effectiveness of search by content. Also creating a validator and converter WSDL to OWL-S, so that the service provider can publish their services without the need to publish its ontology.

At the time of publication happens converting the WSDL to standard OWL-S proposed by Martin et al. [20]. Another possible work is to get the context of social networks, increasing the sensitivity of context.

References

[1] O. F. Ferreira Filho and M. A. G. V. Ferreira, "Semantic web wervices: a restful approach," *Proceedings of the IADIS International Conference on WWW/Internet*, pp. 169–180, 2009.

[2] A. K. Dey, "Providing architectural support for building context-aware applications," Ph.D. dissertation, Georgia Institute of Technology, 2000.

[3] V. V. Dos Santos, "Cemantika: A domain-independent framework for designing context-sensitive systems," Ph.D. dissertation, Universidade Federal de Pernambuco, 2008.

[4] C. E. Cirilo, A. F. Do Prado, W. L. De Souza, and L. A. M. Zaina, "A hybrid approach for adapting web graphical user interfaces to multiple devices using information retrieved from context," in *DMS 2010 - Proceedings of the 16th International Conference on Distributed Multimedia Systems*, 2010, pp. 168–173.

[5] T. Erl, *Soa: principles of service design*. Prentice Hall Upper Saddle River, 2008, vol. 1.

[6] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific american*, vol. 284, no. 5, pp. 28–37, 2001.

[7] F. Curbera, F. Leymann, T. Storey, D. Ferguson, and S. Weerawarana, *Web services platform architecture: SOAP, WSDL, WS-policy, WS-addressing, WS-BPEL, WS-reliable messaging and more*. Prentice Hall PTR Englewood Cliffs, 2005.

[8] H. Haas and A. Brown. (2004) Web services glossary. [Online]. Available: http://www.w3.org/TR/2004/NOTE-ws-gloss-20040211/

[9] Y. L. Nilo Mitra. (2007) Soap version 1.2 part 0: Primer (second edition). [Online]. Available: http://www.w3.org/TR/2007/REC-soap12-part0-20070427/

 ISBN: 978-1-61208-280-6

[10] R. T. Fielding and R. N. Taylor, "Principled design of the modern web architecture," in *Proceedings - International Conference on Software Engineering*, 2000, pp. 407–416.

[11] C. Pautasso, O. Zimmermann, and F. Leymann, "Restful web services vs. "big" web services: Making the right architectural decision," in *Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08*, 2008, pp. 805–814.

[12] W. Jiang, D. Lee, and S. Hu, "Large-scale longitudinal analysis of soap-based and restful web services," in *Proceedings - 2012 IEEE 19th International Conference on Web Services, ICWS 2012*, 2012, pp. 218–225.

[13] M. Hadley. (2009) Web application description language. [Online]. Available: http://www.w3.org/Submission/wadl

[14] S. Vinoski, "Serendipitous reuse," *IEEE Internet Computing*, vol. 12, no. 1, pp. 84–87, 2008.

[15] F. Belqasmi, J. Singh, S. Y. Bani Melhem, and R. H. Glitho, "Soap-based vs. restful web services: A case study for multimedia conferencing," *IEEE Internet Computing*, vol. 16, no. 4, pp. 54–63, 2012.

[16] A. Rodriguez. (2008) Restful web services: The basics. [Online]. Available: http://www.ibm.com/developerworks/webservices/library/ws-restful

[17] F. Belqasmi, R. Glitho, and C. Fu, "Restful web services for service provisioning in next-generation networks: A survey," *IEEE Communications Magazine*, vol. 49, no. 12, pp. 66–73, 2011.

[18] E. Christensen, F. Curbera, G. Meredith, S. Weerawarana *et al.*, "Web services description language (wsdl) 1.1," 2001.

[19] S. Dustdar and W. Schreiner, "A survey on web services composition," *International Journal of Web and Grid Services*, vol. 1, no. 1, pp. 1–30, 2005.

[20] D. Martin, M. Burstein, J. Hobbs, O. Lassila, D. McDermott, S. McIlraith, S. Narayanan, M. Paolucci, B. Parsia, T. Payne *et al.*, "Owl-s: Semantic markup for web services," *W3C member submission*, vol. 22, pp. 2007–04, 2004.

[21] G. Antoniou and F. Van Harmelen, *A semantic web primer*. MIT press, 2004.

[22] M. Forte, W. L. de Souza, and A. F. do Prado, "Using ontologies and web services for content adaptation in ubiquitous computing," *Journal of Systems and Software*, vol. 81, no. 3, pp. 368–381, 2008.

[23] U. Bellur and H. Vadodaria, "Web service ranking using semantic profile information," in *Web Services, 2009. ICWS 2009. IEEE International Conference on*, 2009, pp. 872–879.

[24] P. R. Reddy and A. Damodaram, "Web services discovery based on semantic similarity clustering," in *2012 CSI 6th International Conference on Software Engineering, CONSEG 2012*, 2012.

[25] P. B. Santos, L. K. Wives, and J. P. M. De Oliveira, "An improved approach for measuring similarity among semantic web services," in *WEBIST 2012 - Proceedings of the 8th International Conference on Web Information Systems and Technologies*, 2012, pp. 83–88.

[26] OASIS. (2002) Oasis uddi specificaion tc. [Online]. Available: https://www.oasis-open.org/committees/uddi-spec/faq.php

[27] V. Vieira, P. Tedesco, and A. C. Salgado, "Designing context-sensitive systems: An integrated approach," *Expert Systems with Applications*, vol. 38, no. 2, pp. 1119–1138, 2011.

[28] A. K. Dey, "Understanding and using context," *Personal and ubiquitous computing*, vol. 5, no. 1, pp. 4–7, 2001.

[29] V. Vieira, L. R. Caldas, and A. C. Salgado, "Towards an ubiquitous and context sensitive public transportation system," in *Proceedings - 4th International Conference on Ubi-Media Computing, U-Media 2011*, 2011, pp. 174–179.

[30] R. Gimson, R. Lewis, and S. Sathish, "Delivery context overview for device independence," *W3C Working Group Note*, vol. 20, 2006.

[31] R. S. Pressman, *Engenharia de software*, 6th ed. São Paulo: McGraw-Hill, 2006.

[32] T. Tegegne, B. Kanagwa, and T. van der Weide, "ehealth service discovery framework for a low infrastructure context," in *Computer Technology and Development (ICCTD), 2010 2nd International Conference on*. IEEE, 2010, pp. 606–610.

[33] D. John and M. Rajasree, "A framework for the description, discovery and composition of restful semantic web services," in *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*. ACM, 2012, pp. 88–93.

[34] O. F. Ferreira Filho and M. A. G. V. Ferreira, "Semantic web services: a restful approach," in *Proceedings of the IADIS International Conference on WWW/Internet*, 2009, pp. 169–180.

[35] G. A. Miller. (2013) About wordnet - wordnet. [Online]. Available: http://wordnet.princeton.edu

 ISBN: 978-1-61208-280-6

A New Unsupervised Web Services Classification based on Conceptual Graphs

Eiman Boujarwah
CS Department, Kuwait University
POB 5969 Safat, Kuwait, 13060
ebujarwa@sci.kuniv.edu.kw

Hamdi Yahyaoui
CS Department, Kuwait University
POB 5969 Safat, Kuwait, 13060
hamdi@sci.kuniv.edu.kw

Mohammed A. Almulla
CS Department, Kuwait University
POB 5969 Safat, Kuwait, 13060
almulla@sci.kuniv.edu.kw

***Abstract*— With the drastic growth in number of deployed Web services, the discovery of a desired Web service is becoming a challenging research problem. In this paper, we develop a new unsupervised classification technique of Web services using conceptual graphs. A conceptual graph helps in building functional domains and classifying Web services into these domains. Such classification would speed up the discovery of a Web service and save the time of searching the whole Web service registry. The proposed algorithm is shown to have better performance than the Inductive Reasoning (IR) technique based on OWLS-TC benchmark.**

Keywords-Web services; classification; Conceptual graphs

I. INTRODUCTION

Web services are a predominant information technology for the development of loosely-coupled and cross-enterprise business applications. Hence, Web Services are now considered as a trendy research topic. Several research initiatives investigated Web Services classification for discovery purpose. A Web Service interface is specified using Web Service Description language (WSDL) [13], which is a XML-based description language that describes the functionality of a service. It also provides the parameters that a Web service expects to be input to it along with the output parameters it returns.

The objective of this work is to develop an intelligent technique that leverages conceptual graphs in order to build functional domains and to classify Web services into these domains. Conceptual graphs are considered as a system for knowledge representation based on semantic networks. Conceptual graphs include concepts and relations. This structure should help to bootstrap and speed up the discovery of Web services.

The remaining part of the paper is organized as follows: Section II shows the related work, exploring different methods to classify Web services. Parsing and Stemming the WSDL file is introduced in Section III. Next, Section IV presents the proposed classification of Web Services method using conceptual graphs with the experimental results. Finally, the conclusion and future work are discussed in Section V.

II. RELATED WORK

Classifying Web services is currently an important issue for the Web services community. Several research initiatives tackled this issue from different perspectives. Some of them mainly deal with the description of the Web Service, whereas others focus on the semantic perspective of the interface of a Web service, which is described using WSDL.

Wang et al. [1] proposed a method to manage service classification within a medium or big category. They used Support Vector Machine (SVM) text classification algorithm to classify Web services and they used United Nations Standard Products and Service Code (UNSPSC) as the classification criteria for these Web services.

Meditskos et al. [2] applied several machine learning algorithms to automatically classify Web services into their functional domains based on OWL-S advertisements. They combined the textural description of the Web service and its semantic description. Finally, they compared the accuracy of their algorithm with respect to other algorithms and they found that the semantic signature algorithm achieved better accuracy than the other algorithms.

Segev and Toch [3] provided an analysis of two methods for context-based matching and ranking of Web services for composition purposes. First, they analyzed two common methods for text processing: TF/IDF and context analysis, and two methods of service description namely free text and WSDL. Second, they presented a method for evaluating the proximity of services for possible compositions. Each Web service WSDL context descriptor is evaluated according to its proximity to other services' free text context descriptors. The proposed methods were tested on a large repository of real-world Web services. The experimental results concluded that context analysis is more useful than TF/IDF. Furthermore, the method of evaluating the proximity of the WSDL description to the textual description of other services provides high recall and precision results.

Elgazzar et al. [4] attempted to cluster Web services based on functional similarities. Their proposed technique leveraged the quality threshold of clustering algorithms to cluster similar Web services based on five elements found in the WSDL file namely: WSDL contents, types, messages, ports and service name.

 ISBN: 978-1-61208-280-6

Lately, Kiefer and Briensten [11] came up with a collection of inductive methods to perform classification. Their collection includes Rational Probability Trees (RPT) and Rational Bayes Classifier (RBC). The main idea was to explore links between objects to improve the classification process. Their proposed approach outperformed the kernel methods used in SVM.

Most of the related research initiatives adopt supervised classification methods to classify web services into functional domains. Contrarily, we propose a new unsupervised classification technique based on conceptual graphs. We advocate the use of conceptual graphs to achieve a high level of classification accuracy.

III. WSDL PARSING AND STEMMING

As mentioned in the introduction, the WSDL is a XML-based language, which describes the functionality of the Web service and how to access a particular Web service. A WSDL document consists of four major elements beside the name of the Web service, these elements are: port type, messages, binding and types.

Port type is the most important element in a WSDL document; it describes the operations performed by the Web service. The messages element defines the data elements of an operation. Binding defines the data format and protocol for each operation. Finally, the types element is used as a container for data type definitions in the Web service. Figure 1 shows an example of a WSDL file.

```
<wsdl:portType name="CarPricequalitySoap">
<wsdl:operation name="get_PRICE_QUALITY">
<wsdl:input
message="tns:get_PRICE_QUALITYRequest">
 </wsdl:input>
<wsdl:output
message="tns:get_PRICE_QUALITYResponse">
</wsdl:output>
</wsdl:operation>
</wsdl:portType>
```

Figure 1. An example of a WSDL file

A Web service interface is described in a WSDL document. Our process of parsing a WSDL file [5] starts with extracting the service name, port type, operation names and input/output parameter names from the WSDL file.

After completing the aforementioned parsing, we perform a tokenization step to produce the set of terms for the Web service by filtering the giving names into terms according to several rules such as case changing, use of underscore and hyphenation, and use of numbers. Table I shows a list of examples of the tokenization rules and how it is applied.

TABLE I: TOKENIZATION RULES EXAMPLES

Rule	Original	Tokenized
Case changing	SendEmail	Send, Email
Case changing	getListOfServices	Get, List, of, Services
Underscore	Computer_science	Computer, science
Numbers	Exchange1	Exchange

After that, terms are stemmed into their stemmed version. The stemming process is a well-known process and we use Porter stemmer algorithm [6] to deal with it. Table II shows a list of examples of the stemming process.

TABLE II: STEMMING EXAMPLES

Original	Stemmed
Sending	Send
Cleaned	Clean
Taken	Take
Swimming	Swim
Easier	Easy

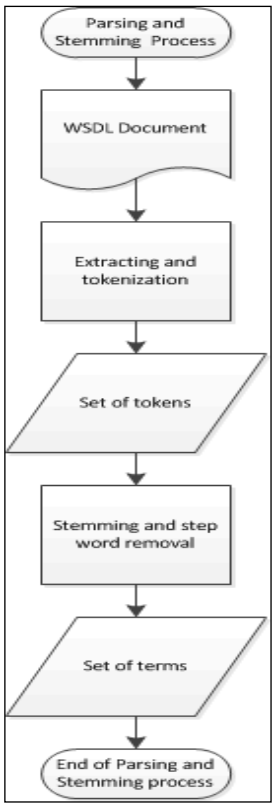


Figure 2. WSDL file parsing process

Finally, we remove the stop-words; these are the words that are most commonly used in any English paragraph, short function words, such as: the, is, at, which and on. These words may cause a problem, so we removed them. The result of this process is a set of terms for each Web service. Figure 2 illustrates the process that we have just described above.

 ISBN: 978-1-61208-280-6

IV. WEB SERVICES CLASSIFICATION

In this research, we consider the problem of how to classify various Web Services into domains according to their functionalities. Domains that will be used in this work are: communication, education, food, travel and weapon. We used the WordNet [7, 8, 9] as a lexical English database, which consists of nouns and verbs; these are grouped into sets of synsets. WordNet is the main conceptual graph and we build the conceptual graph for each domain from it.

We generate a conceptual graph for each functional domain as follows: the terms inside the WSDL are considered as the concepts nodes and the relations are considered as the same semantic relations as the one in WordNet; which are: is-a relations and kind-of relations. A conceptual graph is represented as a hypergraph. A hypergraph is a graph such as each edge can connect to any number of vertices. It is a powerful knowledge representation with higher order relationships between the graph nodes. Figure 3 shows an example of a hypergraph, where V is the set of vertices and E is the set of edges.

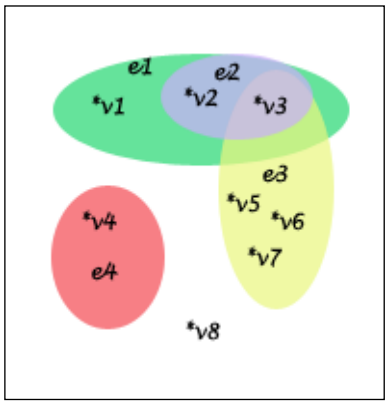


Figure 3. An example of a hypergraph, with $V=\{v_1,v_2,v_3,v_4,v_5,v_6,v_7,v_8\}$

$E=\{e_1,e_2,e_3,e_4\}= \{\{v_1,v_2,v_3\},\{v_2,v_3\},\{v_3,v_5,v_6,v_7\},\{v_4\}\}$

WordNet has a hypergraph that covers all the nouns and verbs (called terms) in the English language along with the semantic relations that connect these terms using the IS-A relation and the KIND-OF relation. We use this hypergraph to be the main conceptual graph from which we generate the hypergraph for each functional domain.

A. HGCA

To start the classification process, we first create a hypergraph for each domain by extracting its contents from the WordNet hypergraph, which are related to a specific domain. We take the name of the domain and start a Breadth First Search in the WordNet hypergraph to create the domain hypergraph. This extraction process will take care of all the connected semantic relations in the new domain's hypergraph. Algorithm 1 shows the steps of generating a domain hypergraph. The implementation was done in the Java programming language and we used the hypergraph API to create, query and search through hypergraphs [10].

After having each domain hypergraph ready, we start the classification process. The hypergraph classification algorithm (HGCA) starts with the file name of the WSDL file for the Web service. First, we open the WSDL file and we perform parsing and stemming on the content of the WSDL file. Then, we compute the classification score for the Web service for each domain. We take the terms of the result from the parsing and stemming step and try to find if this term is matching any of the terms in the domain's hypergraph then we increase the score by one. We repeat these steps for the all terms of a Web Service. After we complete finding the classification score for each domain we compare them all together. The higher the score, the more this Web service belongs to the domain.

Algorithm 1. Generating Domain Hypergraph Algorithm

Algorithm 1: Generating Domain Hypergraph Algorithm
Input: WordNet Hypergraph, Domain Name and K-Level
Output: Domain Hypergraph
begin
Find domain name in WordNet Hypergraph
Start Breadth-First-Search from the domain name until we reach the end of this graph or the K-level.
end
Return
Domain Hypergraph

Our classification algorithm is based on number of nodes matched. We take the Web service and check the number of nodes matched in the domain hypergraph for each domain. For now, we will consider this number as the score for each domain and we will classify the Web service based on the maximum score. Algorithm 2 explains the procedure of the algorithm.

Algorithm 2. HGCA- Hypergraph Classification Algorithm

Algorithm 2:HGCA- Hypergraph Classification Algorithm
Input: WSDL document
Output: Score of the classified functional domain
begin
Filter the WSDL document by parsing and stemming method.
For each Domain
Find matching terms inside the domain hypergraph.
Compute classification score for this domain.
end
Return
The highest score and consider it as the classified domain

B. Experimental Results

To explore the practicality of the proposed technique, a dataset is needed for testing. There are several benchmarks available online among which we chose the (OWLS-TC3)

 ISBN: 978-1-61208-280-6

benchmark [12]. This data set consists of more than 700 Web services, covering seven domains namely: communication, economy, education, food, medicine, travel and weapon. Figure 4 shows the domain norm statistics of the benchmark dataset.

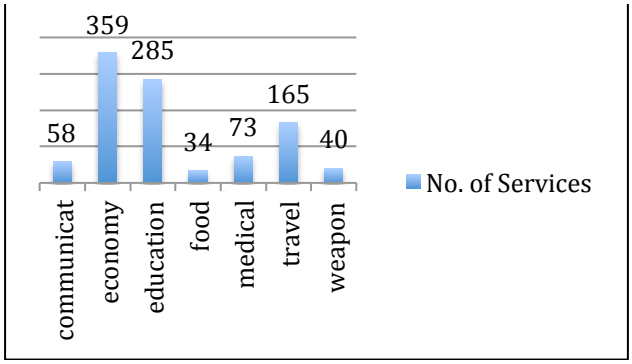


Figure 4. Norm Statistics of OWLS-TC3 benchmark

In our experiments, we randomly selected 200 Web services from the benchmark. It turned out that these Web services belong to five functional domains: communication, education, food, travel and weapon. We got 140 Web services that were correctly classified, 44 Web services not correctly classified and 16 Web services classified not to their functional domains but to other possible domains. For example, a Web service with the name FoodPrice.wsdl should be classified to the economy domain but it appeared in the food domain, which makes sense.

To evaluate the effectiveness of our algorithm, we had to compute the accuracy of the algorithm along with precision and recall. Precision measures the correctness of a classifier and recall measures the completeness of a classifier. When the precision is high, this means that the algorithm has returned more relevant results and when the recall is high it means that the algorithm returned most of the relevant results.

The accuracy of HGCA is almost 80% when we combine the correctly classified Web Service with the Web services that are classified to other possible domains. The precision and recall values for each domain are listed in Table III.

TABLE III: HGCA EFFECTIVNESS

Domain	Comm.	Edu.	Food	Travel	Weapon	Avg.
Precision	0.35	0.61	1	0.95	0.33	0.65
Recall	0.92	0.63	0.94	0.69	0.33	0.70

Finally, we compared our work with the IR method [11]. Table IV shows the average of the precision and recall for our algorithm compared to IR-based algorithm for five functional domains: communication, education, food, travel and weapon.

TABLE IV: PRECISION AND RECALL COMPARISONS

Method	Avg. Precision	Avg. Recall
HGCA	0.65	0.70
IR	0.61	0.42

As can be seen from the table above, the HGCA method outperforms the Inductive Reasoning method in the recall, which means that HGCA returned most of the relevant results and it is more complete. On average, the precision is almost the same with slightly difference between the two methods.

For the same experiment, we computed the average execution time in milliseconds of HGCA; figure 5 shows the scalability of the average execution time in milliseconds of HGCA.

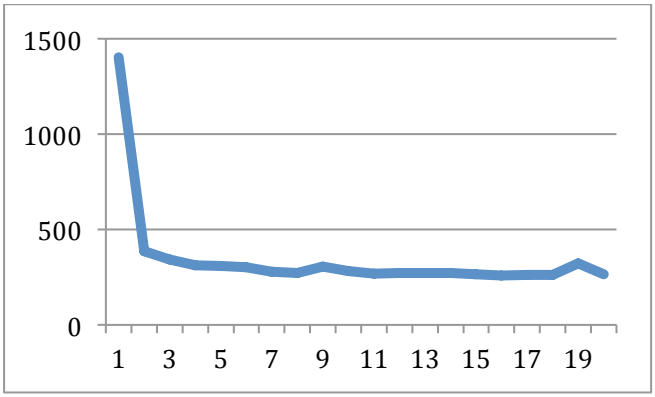


Figure 5. Avarge execution time of HGCA

When the number of services is small it takes longer time because of loading the domain hypergraphs for the first time into the memory.

Figure 6 shows the execution time per domain, we computed the execution time considering only one domain at a time to see the impact of each domain on HGCA.

 ISBN: 978-1-61208-280-6

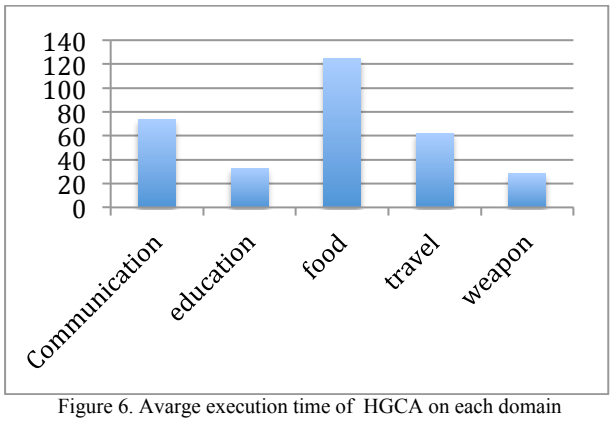


Figure 6. Avarge execution time of HGCA on each domain

The food domain takes longer time than the other domains because of the size of the domain itself to be loaded in the memory. The average execution time of all the domains is 64.4.

V. CONCLUSION AND FUTURE WORK

In this research, we presented a new unsupervised technique for classifying Web services into functional domains based on conceptual graphs. A conceptual graph helps in identifying functional domains and classifying Web services into these domains. Such classification would reduce the search time of specific Web services. In our future work, we are planning to improve our classification method by trying different semantic similarity equations to compute the score of each classified domain; this would give us more accurate results. Furthermore, we are planning to leverage this technique to find the similarity between Web services based on their interfaces that are described in WSDL. The similarity issue between two Web services is equivalent to a matching problem between two conceptual graphs.

REFERENCES

[1] H. Wang, Y. Shi, X. Zhou, Q. Zhou, Sh. Shao, and A. Bouguettaya, "Web Service Classfication using Support Vectore Machine", IEEE International Conference on Tools with Artificial Intelligence, vol. 1, pp. 3-6, 2010.

[2] I. Katakis, G. Meditskos, G. Tsoumakas, N. Bassiliades, and I.P. Vlahavas, "On the Combination of textual and semantic descriptions for automated semantic Web service classification", In AIAI, vol. 296 of IFIP, Springer, pp. 95-104, 2009.

[3] A. Segev, and E. Toch, "Context-Based Matching and Ranking of Web Service for Composition", IEEE Transactions of Services Computing, 2(3): 201-222, 2009.

[4] K. Elgazzar, A. Hassan, and P. Martin, "Clustering WSDL Documents to Bootstrap the Discovery of Web Services", IEEE International Conference on Web Services, Florida, Miami, pp. 147-154, 2010.

[5] E. Boujarwah, "A New Approach to Measuring Similarity Between Web Services", Third Kuwait e-Services and e-Systems Conference (KCESS-2012), Kuwait, 18-20th December, 2012.

[6] K. Sparck Jones, and P. Willet, Readings in Information Retrieval, San Francisco: Morgan Kaufmann, ISBN 1-55860-454-4, 1997.

[7] G.A. Miller, "WordNet: A Lexical Database for English", Communications of the ACM, vol. 38(11): 39-41, 1995.

[8] Ch. Fellbaum, "WordNet: An Electronic Lexical Database", Cambridge, MA: MIT Press, 1998.

[9] Princeton University, "About WordNet", Princeton University, http://wordnet.princeton.edu, 2010.

[10] I. Borislav, K. Vandev, C. Costa, M. Marinov, M. de Queiroz, I. Holsman, A. Picard, and I. Bogdahn, "HypergraphDB 2010", www.hypergraphdb.org, Last visit: 13th October, 2012.

[11] Ch. Kiefer, and A. Bernstein, "Application and Evaluation of Inductive Reasoning Methods for the Semantic Web and Software Analysis", Reasoning Web 2011, LNCS 6848, pp. 460-503, 2011.

[12] M. Klusch and P. Kapahnke, "SemWebCentral OWL-S", http://projects.semwebcentral.org/projects/owls-tc/, Last visit: 21st September, 2010.

[13] E. Christensen,F. Curbera,G. Meredith, and S. Weerawarana, "Web Services Description Language (WSDL)", www.w3.org/TR/wsdl, Last visit: 15th March, 2010.

 ISBN: 978-1-61208-280-6

Enhancing Semantic Web Services Discovery Using Similarity of Contextual Profile

Zahira Chouiref
University Akli Mohand Oulhadj
Bouira, Algeria
Email: zahirachouiref@gmail.com

Abdelkader Belkhir
University of Sciences and Technology
Houari Boumediene, Algiers, Algeria
Email: kaderbelkhir@hotmail.com

Allel Hadjali
Laboratory of Informatique and Automation
Systems - ENSMA, Poitiers, France
Email: allel.hadjali@ensma.fr

***Abstract*—Due to the proliferation of Web services and to the complexity and diversity of users' needs, new efficient mechanisms to automatically discover Web services are strongly required and desired. Recent advances have enabled the supply/consumption of services by suppliers/users of different categories. Every user has some particular interests and preferences when searching appropriate services on the Web. Supporting the "service profile" in the specification of Web services becomes then a paramount important issue. It can especially help applicants and Web services providers to produce accurate descriptions and thus improve the degree of relevance of the responses returned in the process of Web services discovery. In this paper, we present a novel method in the view of users requirements for Web services discovery. Especially, we propose (i) a framework to model the end user and the Web service contextual profile for best search, (ii) as well as a solution of semantic Web services discovery based on contextual profile similarity. This similarity is performed using a hybrid similarity measure we have defined.**

***Keywords*—Semantic Web Services; Context; Profile; Semantic Web Services Discovery; Hybrid Similarity Measure.**

I. INTRODUCTION

The evolution of the IT (Information Technology) tools led to the development of new paradigms that describe interactions that exist over IT applications such as SOA (Service-Oriented Architecture) [15]. Service orientation is a promising paradigm for offering and consuming functionalities within and across organizations in the Web. Indeed, SWS (Semantic Web Services) allow a homogeneous use of heterogeneous software components deployed in large networks and in particular the Internet.

Several studies have already been done in the area of SWS discovery which is the process permitting to find the most suitable set of Web services that fulfill some functionality. Some of the discovery approaches are syntax based [3][5], while other are semantic based [7][8][9]. Despite the simplicity and facility of their implementation, syntactic approaches exhibit some serious limitations. The predominant problem is the restrictions posed by keywords matching that do not allow retrieval of SWS with similar functionalities; two WSDL (Web Services Description Language) descriptions can be used to describe the same service but with different words. However, when modeling Web services using ontologies, the semantic representation of concepts and their relations can be exploited and thus semantic matching to be performed. To enable SWS discovery, the first thing we need is a semantic service description language OWL-S (Ontology Web Language for Services) [14]. This development is significant since it seems to be able to approach certain insufficiencies of the syntactic approaches and tackle some of the UDDI (Universal Description, Discovery and Integration) register inadequacies.

However, with the exponential growth of SWS [17], the diversity of users and the conditions under which they access Web services, finding the relevant SWS that fit the needs and the context of the user is becoming a challenging task. With Web 2.0 applications and particularly e-business and e-commerce applications, SWS discovery is becoming much more important in a Web context. Our approach aims at solving queries using the profile of users from a contextual informational view of the Web where users have several features such as the client terminal, the client preferences, his/her location, etc. All these parameters form a particular context of user called the *contextual profile*.

The notions of context and profile have been the subjects of many works. Several meanings of such notions exist in the literature. For instance, and as pointed out in [21], "*A profile is defined by a set of attributes, possibly organized into abstract entities, whose values can be user-defined or dynamically derived from user behavior. A profile is supposed to characterize user domain of interest and all his specific features that help the information system to deliver the most relevant data in the right form at the right place and the right moment*". Since generally a person has several interests, user profiles should be defined to represent the various interests of the user. In [22], a user profile is proposed to represent the distinct interests related to a user. Various definitions of the context are also given and summarized in [10]. Brown et al. [23] define the context as being information about *location*, *the identity of people in close proximity*, *physical conditions*. Ryan et al. [24] add to this definition the notion of *time*. In [25], four categories of context {*location, identity (user), activity (state), time*} are identified as the more important parameters in practice. As stated previously, it is not easy to give a complete definition for a context. In fact, the notion of context is not universal but relative to some situation and application domain [23].

Challenges. The problem of interest is that the search engine produces several results in response to a user query,

 ISBN: 978-1-61208-280-6

with the consequence that the truly relevant results are often missed.

- In order to face this unsatisfactory situation, one of the suggested solutions leverages information related to user profile. Indeed, the results produced by a search engine for a given query are not the same, dependent on the user profile and the context in which the user made the request. The main idea is to understand what the profile information is depending on a user context during a session: How do results depend on profile and change in context? The service context can group the *service localization (geographical restriction)*, *the implementation cost*, *the QoS (Quality of Service) parameters*, etc. The user context can be formed of his/her *localization*, his/her *devices*, etc. Since the user profile can vary during a session, the system must be able to adapt it in order to select services according to the new context. "*A contextual profile of an end user consists of a list of concepts which specifies the user domain of interest, personal information (age, sex, etc.), data quality, data on behavior, preferences and security, devices, purpose in a given temporal context, social context, spatial context and informative context*". Personal characteristics can strongly influence the interaction with a system and security can distinguish the user from the others, in a given context. Suppose a doctor is looking for a hotel that is in close proximity to the airport, with shuttle service, and with at least one conference room that has Internet connection and a 3D printer. The doctor is performing his query via the smartphone. Current search engines provide a list of all the hotels, but the question remains: which one to choose? End user context consists of his/her device (smartphone) from which we can deduce its navigator, operating system, location and the type of printer that will be used (wifi printer), etc. The end user is considered as guest of the secured system; therefore, the access to the printer and Internet is controlled. According to the user profile (guest or administrator), the user should have received a confidential message regarding the access rights to the resources. This is done to identify the sender so that the system can determine the user's access rights. His/Her profile consists of his/her profession (doctor) and his/her preferences (hotel with conference room, Internet, 3D printer, close to the airport and shuttle service, etc).
- In addition, the methods available in the UDDI publication do not contain a formal model describing the context of services use. Therefore, services discovery could not be achieved efficiently without considering their contextual profile. When a user requests a Web services discovery system, (s)he would have services tailored to his/her context and profile.

Contributions. Our paper specifically includes the following contributions:

- Combining context and profile models in one model called *Contextual Profile Model* that takes advantages of both models. The proposed model provides relevant and adapted results to the user requirements. It also allows the representation of all information that characterizes both the user and the service.
- We allow discovering suitable SWS based on the calculation of similarity between the user profile and the service profile according to a given context (the same query issued by different users may have different results as it is evaluated using different profiles in different contexts).

In the remainder of this paper, we first define in Section II an overview of existing efforts towards the SWS discovery and then we provide a critical analysis of these approaches. In Section III, we set up the contextual profile model firstly, so as to present a formalism of service and user in details; then, we present SWS discovery architecture. Finally, Section IV depicts a conclusion with the main research directions.

II. Related Word and Discussion

Our work can be positioned in the new interdisciplinary area of Web Science which is the effort to bridge and formalize the semantic and technical aspects of the World Wide Web.

Previous works [7][8][9] have focused mainly on providing means to describe the functionality of a SWS and to allow a very expressive language for querying services. However, none of the works discusses in depth the concept of the profile and how a publisher/requester should provide context data about his/her services. The information should have more user context centric presentation in the discovery system. For instance, multiple Web services with similar functionalities are often available, best service(s) among them should be selected. To achieve this goal, one way is to use context [20] and QoS parameters which, generally, include performance, usability, safety, cost, etc.

The studies in [13] focus on QoS in discovery systems. The service consumer searches UDDI registry for a specific service through discovery agent which helps to find best quality service from available services which satisfies QoS constraints and preferences of requesters. Context-Awareness as proposed in [11] performs the necessary changes in the service behavior and/or the data handled in order to adapt the service to the context of the each user. Rong et al. [12] suggest with an example that context should be domain oriented or problem oriented in Web services discovery system. They divide context in two categories as explicit and implicit, with Personal profile oriented context, Usage history oriented context, Process oriented context and other context. Chukmol et al. [6] propose the personal opinion on service functionality and quality or invocation cost should also be considered by collaborative tagging-based environment for Web services discovery. The study done in [4] has proposed a novel approach to enhance Web services discovery based on, among others, QoS, customer's preferences and past experiences. The work in [26] presents an alternative approach for supporting users in Web services discovery by implementing the implicit culture approach for recommending Web services to developers based

 ISBN: 978-1-61208-280-6

on the history of decisions made by other developers with similar needs.

The limitation of these approaches is that they make the system architecture more complicated when new attributes and constraints are introduced. SWS properties include several parameters like the functional (*Input, Output, Preconditions, Effects, functionalities,* etc.) and non-functional parameters (*QoS, property identifies the technical standards or protocols for implementing services*, and *categorization*). However, the majority of suggested approaches focus only on some parameters: *QoS, localization, user behavior.* Moreover, few works took into account multiples qualitative and quantitative parameters to help users to find the best service during the discovery process. It is also noticed that the suggested semantic approaches are based on the same technique, which consists in calculating the semantic correspondence level between the functional and non-functional parameters of the services and those cited in the user request. One of the big problems of Web search is the definition of a correspondence function between the representation of the proposed service and the user request. In order to fill this gap, we propose a new approach to improve the automatic SWS discovery, based on the measure of similarity between the user contextual profile and the one of the available services.

Thus, Web services that fit better the profile and the context of end user are retrieved. In our approach, we also indicate the interest of the proposed similarity measure in order to sort the candidate services due to the profile and the context attributes.

III. A Contextual Profile Similarity Based Semantic Web Services Discovery Approach

We present in this section the service, the user and the contextual profile formalism, as well as the proposed architecture. Finally, we show how a new similarity measure can be used to enhance the discovery process.

A. Model of Service/User Contextual Profile

Different attempts have been done to collect and classify profile's information. Most of the profile categorization has been done by Amato et al. [2].

The information of the contextual profile can be static (personal data, etc), evolutionary (intellectual quality, preferences, etc) and temporary (localization, devices, etc). These pieces of information must be captured to match demands to offers of services, on the syntactic and semantic level in order to improve the relevance of answers during a discovery session. In Figure 1, we show the proposed model that contains several dimensions able to describe the most information characterizing a profile. This general structure takes the form of a tree that contains a hierarchy of concepts. Each concept is constituted of one or several sub-concepts, that contain to their turn one or several attributes. The structure thus defined is flexible in the sense that different features can be spread through the tree structure of the proposed description. It permits to model the user's contextual profile soliciting the service as well the SWS's contextual profile offered.

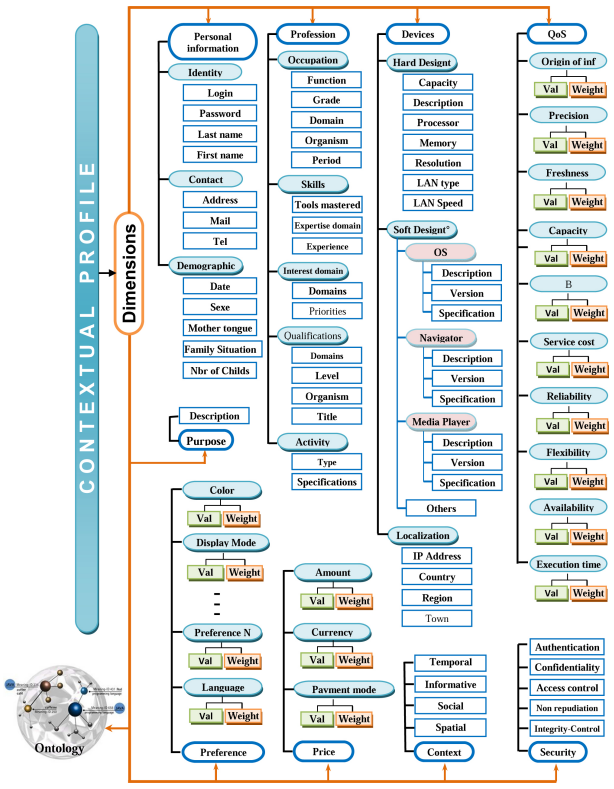


Fig. 1. Concepts and sub-concepts of the contextual profile of the service/user.

B. Proposed Discovery Architecture

Our approach supported by the architecture depicted in Figure 2, is capable of integrating the contextual profile into the process of SWS discovery. It is composed of: *(i) basic elements of SOA* [15] namely: the service requester, the service provider and the service registry; *(ii) Interoperability Module* which comprises an administrator of the profile, an inter ontology similarity module and a contextual profile database, *(iii) Discovery and Selection Engine* which contains a contextual profile filtering module of Service/Request and a similarity measure treatment module.

C. Semantic Web Service, Request and Contextual Profile Formalism

Formally, an SWS is defined as a quintuple $\{n_s, d_s, p, op, cpp\}$ such that:

- n_s is the name of the SWS.
- d_s is the functional description of the SWS.
- p is the set of parameters describing the SWS.
- op is the set of operations of the SWS.
- cpp is the set of concepts constituting the contextual profile of provided service.

Operation Op is defined as a quintuple $\{n_o, d_o, i, o, pre, ef\}$ such that:

- n_o is the name of the operation.
- d_o is the functional description of the operation.
- $i = (i_1, \ldots i_n)$ is the set of input parameters of the operation

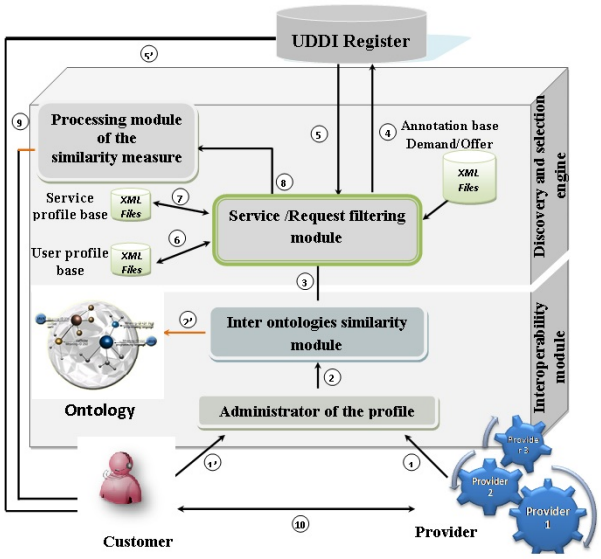


Fig. 2. Proposed architecture.

- $o = (o_1, \ldots o_n)$ is the set of output parameters of the operation.
- prc specifies the preconditions of the service.
- ef specifies the effects of the service.

Request R is a quintuple $\{n_r, i_r, o_r, cpr, s\}$ such that:

- n_r is the name of the request.
- $i_r = (i_{r1}, \ldots i_{rn})$ is the set of input parameters of the request.
- $o_r = (o_{r1}, \ldots o_{rn})$ is the set of output parameters of the request.
- cpr is the set of the concepts constituting the contextual profile of the required service (user contextual profile).
- s $(0 < s < 1)$ is the interest score defined by the user.

Each parameter (*i*, *o*) in the list of (inputs, outputs) respectively is defined as a quadruple $\{n_t, v, t, c_o\}$ such that:

- n_t is the name of the attribute.
- v is the value of the attribute.
- t is the data type of the attribute value.
- c_o is the concept of the ontology which is connected to the attribute to improve its semantic quality. Some approaches use ontologies for the purpose of searching the context and the interests of the users [18][19].

Contextual profile is modeled through a set of concepts, noted CP_i (contextual profile parameters). Each CP_i is characterized by a set of values and weight associated with each parameter, $CP_i = (v_i, w_i)$.

D. A New Similarity Measure

As we described above, the profile is modeled as multidimensional attributes through the tree structure of the description proposed in Figure 1. In general, there are two different approaches for expressing these attributes: a quantitative and a qualitative one. In our work, we propose to improve the similarity measure defined in [16] by combining syntactic with semantic similarity of each qualitative/quantitative attribute of request/service contextual profile. This hybrid similarity measure allows us to compute an overall score for each service retrieved and thus to recommend to the user the best services that match his/her profile. The similarity measure introduced can be formalized as follows:

- Let P be a set of objects' profile (users, documents, Web services, etc.). An object is described by m contextual characteristics $X = (CP_1, CP_2, \ldots CP_m)$.
- Let N be a set of services. Each service is composed of a set of contextual profile concept.
- Let CP_x, CP_y be a contextual profile belonging to P for a given request and service respectively, and $(w_1, \ldots w_m)$ is a set of weights associated to each characteristic of CP_x, where $\sum w_i = 1$ (w_i is a real number between 0 and 1).
- We define a threshold s in order to present only services that have a rate of similarity with the user profile higher than the defined threshold .

The quantitative similarity measure $QSim$ [16] is defined as follows:
$QSim : P \times P \rightarrow [0, 1]$

$$QSim(x, y) = \frac{a \sum_{i=1}^{a} w_i \times sima(x_i, y_i)}{(a \sum_{i=1}^{a} w_i) + (b \sum_{i=1}^{b} w_{a+i}) + (c \sum_{i=1}^{c} w_{a+b+i})} \quad (1)$$

where x_i and y_i are numerical values.

As for our similarity measure $SimPro$, it is given by the following formula:
$SimPro : P \times P \rightarrow [0, 1]$

$$SimPro(CPx, CPy) = \frac{a}{a+b} \times \frac{a \sum_{i=1}^{i=a} w_i \times sima(CPx_i, CPy_i)}{(a \sum_{i=1}^{i=a} w_i) + (b \sum_{i=1}^{i=b} w_{a+i})} \quad (2)$$

- a is the set of common characteristics of CP_x (requested by the client) and CP_y (suggested by the service).
- b is the set of characteristics existing in CP_x and not existing in CP_y (requested by the client but not suggested by the service).
- c is the set of characteristics existing in CP_y and not existing in CP_x (suggested by the service but not requested by the client).
- sima is the atomic similarity between each characteristic of CP_x and CP_y. It is defined in a universe **U** and can be modeled as follows:

$sima : \mathbf{U} \times \mathbf{U} \rightarrow [0, 1]$

$$sima(CPx_i, CPy_i) = \begin{cases} 1 & if \quad c_1 \\ 0 & if \quad c_2 \\ \frac{min(CPx_i, CPy_i)}{max(CPx_i, CPy_i)} & if \quad c_3 \\ value \in [0, 1] & if \quad c_4 \end{cases} \quad (3)$$

With:

$c_1 \Leftrightarrow CPx_i = CPy_i$.
$c_2 \Leftrightarrow CPx_i = 0 \vee CPy_i = 0 \vee CPx_i = \perp \vee CPy_i = \perp$.
$c_3 \Leftrightarrow CPx_i, CPy_i$ are quantitative values.
$c_4 \Leftrightarrow CPx_i, CPy_i$ are qualitative values.

 ISBN: 978-1-61208-280-6

if CPx_i and CPy_i are qualitative values, *sima* is calculated (in our example) by using the Jaccard coefficient [1].

$SimPro$ checks the properties of similarity measures defined in [16]. Let us note also that the most powerful strength of the quantitative/qualitative similarity measure proposed is its ability to take into account attributes of different natures (either qualitative or quantitative) both in the contextual profile of the request and the contextual profile of the Web services. Other approaches borrowed from information retrieval or domain ontology [27], can be used for measuring atomic similarity ($sima$) between concepts.

Note also that the performance of Web service may depend on the number of features it was mentioned by the user and published by the provider, the weighted atomic similarity of each request's characteristic and the quantity (a/a+b) that is in fact the average frequency of all mentioned attributes.

IV. Conclusion and Future Work

We have proposed in this work a new approach to SWS discovery. The aim of this approach is to automatically discover relevant services based on measuring the similarity of user request and SWS using the contextual profile information during the search and selection steps. A quantitative and qualitative similarity measure is applied for the management of the contextual profiles. This hybrid similarity allows to retrieve SWS that better satisfy the user needs. We plan to conduct thorough experiments to study the effectiveness of the proposed formula for measuring similarity, to analyze the quality of SWS from a user point of view, and to consider the use of ontology for context and profile concepts. Another idea to improve the quality of the answers is to consider the parameters of user profile modeled using gradual concepts which can be represented thanks to fuzzy sets.

References

[1] R. Real, "Tables of Significant Values of Jaccard's Index of Similarity," Misc. Zool, 22.1, 1999, pp. 29-40.

[2] G. Amato, U. Straccia, "User Profile Modeling and Applications to Digital Libraries," in Proceedings of the Third European Conference on Research and Advanced Technology for Digital Libraries, Paris, France, 1999, pp. 184-197.

[3] M. B. Blake and M. F. Nowlan, "A Web Service Recommender System Using Enhanced Syntactical Matching," IEEE International Conference in Web Services ICWS, USA, January.2007, pp.575-582.

[4] R. Benaboud, R. Maamri, and Z. Sahnoun, "User's Preferences and Experiences based Web Service Discovery using Ontologies," in Proceedings of the Fourth International Conference on Research Challenges in Information Science (RCIS), Nice, France, May.2010, pp. 121-126.

[5] P. Palathingal and S. Chandra, "Agent Approach for Service Discovery and Utilization," in Proceedings of the 37th Annual Hawaii Internationnal Conference on System Science HICSS-37, USA, January.2004.

[6] U. Chukmol, A. Benharkat, and Y. Amghar, "Enhancing Web Service Discovery by using Collaborative Tagging System," IEEE 4th International Conference on Next Generation Web Services Practices, October.2008, pp. 54-59.

[7] F. Nawaz, K. Qadir, and H. Farooq Ahmad, "SEMREG-Pro: A Semantic based Registry for Proactive Web Service Discovery using Publish Subscribe Model," Fourth International Conference on Semantics, Knowledge and Grid, IEEE Xplore, Pakistan, December.2008, pp. 301-308.

[8] C. Wu, E. Chang, and A. Aitken, "An Empirical Approach for Semantic Web Services Discovery," 19th Australian Conference on Software Engineering, IEEEXplore, Perth, March.2008, pp. 412-421.

[9] G. Wen-yue, Q. Hai-cheng, and C. Hong, "Semantic Web Service Discovery Algorithm and its Application on the Intelligent Automotive Manufacturing System," International Conference on Information Management and Engineering, IEEEXplore, China, April.2010, pp. 601-604.

[10] M. Bazire and P. Brezillon, "Understanding Context before using it," in CONTEXT'05 Proceedings of 5th International Conference on Modeling and Using Context, 2005, pp. 29-40.

[11] M. Keidl and A.Kemper, "Toward Context/aware Adaptable Web Services," in 13th international World Wide Web Conference, USA, 2004, pp. 55-65.

[12] W. Rong and K. Liu, "A Survey of Context Aware Web Service Discovery: From User's Perspective," Fifth IEEE International Symposium on Service Oriented System Engineering, June.2010, pp. 15-22.

[13] T. Rajendran and P. Balasubramanie, "An Optimal Agent-Based Architecture for Dynamic Web Service Discovery with QoS," Second International Conference on Computing Communication and Networking Technologies, IEEEXplore, July.2010, pp. 1-7.

[14] D. Martin and all, "OWL-S: Semantic Markup for Web Services," W3C Member Submission (2004) http://www.w3.org/Submission/OWL-S, November.2004 [retrieved:April, 2013].

[15] D. Booth, and all, "Web Services Architecture," http ://www.w3.org/TR/2003/WD-ws-arch-20030808/, Aout.2003 [retrieved:April, 2013].

[16] A. Belkhirat, A. Belkhir, and A. Bouras, "A New Similarity Measure for the Profiles Management," in the Proceedings of the 13th International Conference on Computer Modeling and Simulation (UKSIM), Cambridgeshire United Kingdom, April.2011, pp. 255-259.

[17] T. Berners-Lee, J. Hendler, and O. Lassila, "The Semantic Web," Scientific American, vol. 284, no. 5, May.2001, pp. 34-43.

[18] C. Bizer, and all, "A Crystallization Point for the Web of Data," Web Semantics : Science, Services and Agents on the World Wide Web, vol. 7, n 3, September.2009, pp. 154-165.

[19] S. Calegari and G. Pasi, "Ontology-Based Information Behaviour to Improve Web Search," Future Internet, vol. 2, n 4, October.2010 pp. 533-558.

[20] M. Baldauf, S. Dustdar, and F. Rosenberg, "A Survey on Context-aware Systems," International Journal of Ad Hoc and Ubiquitous Computing , vol2 (4), June.2007, pp. 263-277.

[21] M. Bouzeghoub and D. Kostadinov, "Data Personalization: a Taxonomy of User Profiles Knowledge and a Profile Management Tool," Reports of the laboratory PRISM, http://www.researchgate.net/publication/228768551_Data_Personalization_a_Taxonomy_of_User_Profiles_Knowledge_and_a_Profile_Management_Tool/file/9fcfd50cf300649190.pdf, 2007 [retrieved:April, 2013].

[22] G. Bordogna and G. Pasi, "A Flexible Multi Criteria Information Filtering Model," Soft Comput. DOI: 10.1007/s00500-009-0476-3. January.2010, pp. 799-809 [retrieved:April, 2013].

[23] P. J. Brown, J. D. Bovey, and X. Chen, "Context-aware Applications: From the Laboratory to the Marketplace," IEEE Personal Communications, October.1997, pp. 58-64.

[24] N. Ryan, J. Pascoe, and D. Morse, "Enhanced Reality Fieldwork: the Context Aware Archaeological Assistant," Dingwall, L., S. Exon, V. Gaffney, S. Laflin and M. van Leusen (eds.), Archaeology in the Age of the Internet. CAA97. Computer Applications and Quantitative Methods in Archaeology. Proceedings of the 25th Anniversary Conference, University of Birmingham, (BAR International Series 750). Archaeopress, Oxford, April.1997, pp. 269-274.

[25] A. K. Dey, G. D. Abowd, and D. Salber, "A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-aware Applications," Human-Computer Interaction Journal 16(2), December.2001, pp. 97-166.

[26] N. Kokash, A. Birukou, and V. D'Andrea, "Web Service Discovery Based on Past User Experience," in Proceedings of the International1 Conference on Business Information Systems (BIS), LNCS Volume 4439, 2007, pp. 95-107.

[27] Y. Tsai, S. Hwang, and Y. Tang, "A Hybrid Approach to Automatic Web Services Discovery," in International Joint Conference on Service Sciences, IEEEXplore, 2011, pp. 277-281.

 ISBN: 978-1-61208-280-6

Internet of Threads

Renzo Davoli
Computer Science and Engineering Department
University of Bologna
Bologna, Italy
Email: renzo@cs.unibo.it

***Abstract*—In the beginning, Internet and TCP/IP protocols were based on the idea of connecting computers, so the addressable entities were networking adapters. Due to the evolution of networking and Internet services, physical computers no longer have a central role. Addressing networking adapters as if they were the true ends of communication has become obsolete. Within Internet of Threads, processes can be autonomous nodes of the Internet, i.e., can have their own IP addresses, routing and QoS policies, etc. In other words, the Internet of Threads definition enables networked software appliances to be implemented, These appliances are processes able to autonomously interoperate on the network, i.e., the software counterpart of the Internet of Things objects. This paper will examine some usage cases of Internet of Threads, discussing the specific improvements provided by the new networking support. The implementation of the Internet of Threads used in the experiments is based on Virtual Distributed Ethernet (VDE), Contiki and View-OS. All the software presented in this paper has been released under free software licenses and is available for independent testing and evaluation.**

***Keywords*-Internet; IP networks; Virtual Machine Monitors**

I. INTRODUCTION

The Internet was designed to connect computers or, more precisely, networking controllers. In fact, IP addresses were assigned, and most of the time are still assigned, to the hardware interfaces [1].

In a typical situation when a client application wants to connect to a server daemon it first uses a Domain Name Server (DNS) to resolve a logical name of the server to an IP address. Usually the DNS maps the logical name, which is a readable specification of the required service (e.g., www.whitehouse.gov or ftp.ai.mit.edu), to the IP address of a hardware network controller of a computer able to provide the requested service.

By Internet of Threads (IoTh) we mean the ability of processes to be addressable as nodes of the Internet, i.e., in IoTh processes play the same role as computers, being IP endpoints. They can have their own IP addresses, routing and QoS policies, etc.

On IPv4, IoTh usage can be limited by the small number of available IP addresses overall, but IoTh can reveal all its potential in IPv6, whose 128-bit long addresses are enough to give each process running on a computer its own address.

This change of perspective reflects the current common perception of the Internet itself. Originally, Internet was designed to connect remote computers using services like remote shells or file transfers. Today most of the time users are mainly interested in specific networking services, no matter which computer is providing them. So, in the early days of the Internet, assigning IP addresses to the networking controllers of computers was the norm, while today the addressable entity of the Internet should be the process which provides the requested service.

For a better explanation, let us compare the Internet to a telephone system. The original design of the Internet in this metaphor corresponds to a fixed line service. When portable phones were not available, the only way to reach a friend was to guess where he/she could be and try to call the nearest line. Telephone numbers were assigned to places, not to people. Today, using portable phones, it is simpler to contact somebody, as the phone number has been assigned to a portable device, which generally corresponds to a specific person.

In the architecture of modern Internet services there are already exceptions to the rule of assigning IP addresses to physical network controllers.

- Virtual Machines (VM) have virtual network controllers, and each virtual controller has its own IP address (or addresses).
- Each interface can be assigned several IP service oriented addresses. For example, if a DNS maps www.mynet.org to 1.2.3.4, and ftp.mynet.org to 1.2.3.5, it is possible to assign both addresses to the same controller. Services can be assigned to a specific process using the *bind* system call.
- Linux Containers (LXC), as well as Solaris Zones [2], [3], allow system administrators to create different operating environments for processes running on the same operating system kernel. Among the other configurable entities for containers, it is possible to define a specific network support, and to create virtual interfaces of each container (flag CLONE_NEWNET of clone(2)). The definition and configuration of network containers, or zones, are privileged operations for system administrators only.

The paper will develop as follows: Section II introduces the design and implementation of Ioth, followed by a discussion in Section III. Related work is described in Section IV. Section V is about usage cases. Section VI discusses the security issues reated to IoTh and Section VII provides some performance figures of a proof-of-concept implementation. The paper ends with some final considerations about future work.

 ISBN: 978-1-61208-280-6

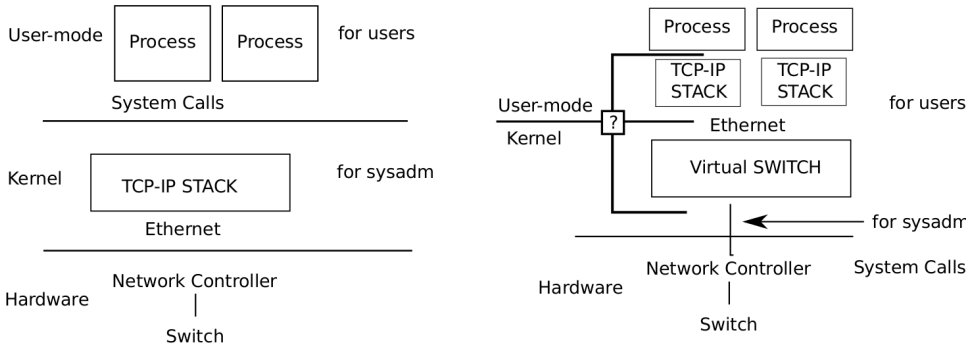


Fig. 1. Different perspectives on the networking support: the standard OS support is on the left side, IoTh is on the right side

II. Design and Implementation of IoTh

The role and the operating system support of the Data-Link networking layer must be redesigned for IoTh. Processes cannot be plugged to physical networking hubs or switches as they do not have hardware controllers (In the following the term *switch* will be used to reference either a switch or a hub. as the difference is not rlevant to the discussion). On the other hand, it is possible to provide processes with virtual networking controllers and to connect these controllers to virtual switches. Figure 1 depicts the different perspectives on the networking support. The focus of Fig. 1 is to show how IoTh changes the Operating System (OS) support for networking: what is provided by the hardware vs. what is implemented in software, what is shared throughout the system vs what is process specific and what is implemented as kernel code vs. what runs in user mode.

The typical networking support is represented on the left side of Fig.1. Each process uses a networking Application Program Interface (API), usually the Berkeley sockets API [4], to access the services provided by a single shared stack, or by one of the available stacks for zones or LXC (see Section I). The TCP-IP stack is implemented in the kernel and directly communicates with the data-link layer to exchange packets using the physical LAN controllers.

In IoTh, represented on the right side of the Figure, unprivileged processes can send data-link packets using virtual switches, able to dispatch data-link packets from process to process and between processes and virtual interfaces (e.g., tuntap interfaces) of the hosting OS. Virtual switches can also be interfaced to physical networking controllers, but this latter operation is privileged and requires specific capabilities (CAP_NET_ADMIN).

So, the hardware-software boundary has been moved downwards in the IoTh design. In fact, the data-link networking (commonly the Ethernet) includes software components in IoTh, i.e., virtual switches, for unprivileged user processes. In IoTh the virtual switches are shared components between user processes, while the TCP-IP stacks (or, in general, the upper part of the networking stack, from the networking layer up) are process specific. It is also possible for a group of processes to share one TCP-IP stack, but in the IoTh design this is just an implementation choice and no longer an OS design issue or system administration choice.

The kernel/user-mode code boundary is flexible in IoTh: both the virtual ethernet switches and the TCP-IP stacks can be implemented in the kernel or not. A virtual switch can be a standard user-mode process or a kernel service, while a TCP-IP stack is a library that can be implemented in the kernel to increase its performance.

III. Discussion

All the concepts currently used in Local Area Neworking can be applied to IoTh networking.

Virtual switches define virtual Ethernets. Virtual Ethernets can be bridged with Physical Ethernets, so that workstations, personal computers or processes running as IoTh nodes are indistinguishable as endnodes of Internet communication. Virtual Ethernets can be interconnected by Virtual Routers. It is possible to use DHCP to assign IP addresses to processes, to use IPv6 stateless autoconfiguration, to route packets using NAT, to implement packet filtering gateways, etc.

IoTh can support the idea of network structure consolidation in the same way that the Virtual Machines provided the idea of Server consolidation. Complex networking topologies can be virtualized, thus reducing the costs and failure rates of a hardware infrastructure. IoTh adds one more dimension to this consolidation process: it is possible by IoTh to virtualize not only each server as such, but also the pre-existing network infrastructure.

Network consolidation is just an example of IoTh as a tool for compatibility with the past. In this example each process joining the virtual networks is just a virtual machine or a virtual router or firewall. The granularity of an Internet node is flexible in IoTh. A virtual machine can be an Internet node, but each browser, bit-torrent tracker, web server or mail transport agent (MTA) can be an Internet node, too.

By IoTh, TCP-IP networking also becomes an Inter Process Communication (IPC) service. A process can have its own IP address(es) and can interoperate with other processes using standard protocols and standard ports. Several processes running on the same host can use the same port, since each one uses different IP addresses. The same IPC protocols can be used regardless of the host on which the process is running:

 ISBN: 978-1-61208-280-6

nothing changes, whether the communicating processes are running on the same host or on different, perhaps remote, computers. This allows a simpler migration of services from one machine to another.

Each process in IoTh can use Internet protocols as its user interface. This means, for example, that it is possible to create programs which register their IP addresses in a dynamic DNS, and which have their web user interface accessible through a standard browser.

IoTh is, from this perspective, the software counterpart of Internet of Things (IoT [5]). In IoT hardware gadgets are directly connected to the network. They interact between objects and with the users through standard Internet protocols. IoTh applies the same concept to processes, i.e., to software objects as if they were virtual IoT gadgets. These IoTh-enabled processes using internet protocols to interoperate can be called networked virtual appliances. If they were implemented on specific dedicated hardware objects, they would become things according to the definition of IoT.

IV. Related Work

IoTh uses and integrates several concepts and tools already available in the literature and in free software repositories.

A. Virtual Ethernet Services

IoTh is based on the availability of virtual data-link layer networking services, usually virtual Ethernet services, as Ethernet is the most common data-link standard used.

The idea of a general purpose virtual Ethernet switch for virtual machines has been implemented by some projects:

- VDE [6] is a general purpose, distributed support for virtual networking. A vde_switch is the virtualized counterpart of an Ethernet switch. virtual machines (or virtual appliances) can be connected to a vde_switch using the vde_plug library. Remote vde_switches can be connected together to form extended LANs. VDE is a service for users: the activation of a VDE switch, the connection of a VM to a switch, or the interconnection of remote switches, are all unprivileged operations. VDE provides support for VLANs, fast spanning trees for link fault tolerance, remote management of switches, etc.
- OpenVswitch [7] is a virtual Ethernet switch for VMs implemented at kernel level. OpenVswitch has VLAN and QoS support. It has been designed to be a fast, flexible support for virtual machines running on the same host. It does not support distributed virtual networks, and requires root access for its configuration.
- Vale [8] is a very fast support for virtual networking, based on the netmap [9] API. It uses shared memory techniques to speed-up the communication between the VMs. Vale, like OpenVswitch, does not directly support distributed networks and must be managed by system administrators.

B. TCP-IP stacks

As described in the introduction, the TCP-IP networking stack is generally unique in a system and it is considered as a shared systemwide service provided by the kernel. Adam Dunkels wrote two general purpose and free licensed TCP-IP stacks for embedded systems: uIP [10] and LWIP (Light Weight IP) [11]. uIP is a very compact stack for microcontrollers having limited resources, while LWIP is a more complete implementation for powerful embedded machines. LWIP was initially designed for IPv4, but a basic support for IPv6 has recently been added. In 2005, when LWIP did not support IPv6 yet, VirtualSquare labs created a fork of LWIP named LWIPv6 [12]. LWIPv6 then evolved independently and is now a library supporting both IPv4 and IPv6 as a single hybrid stack, i.e., differently from the dual-stack approach, LWIPv6 manages IPv4 packets as a subcase of IPv6 packets. When LWIPv6 dispatches an IPv4 packet it creates a temporary IPv6 header, used by the stack, which is deleted when the packet is delivered. LWIPv6 is also able to support several concurrent TCP-IP stacks. It has features like packet filtering, NAT (both NATv4 and NATv6), slirp (for IPv4 and IPv6).

C. Process/OS interface

In this work we use two different approaches to interface user processes with virtual stacks and virtual networks. A way to create networked software appliances is to run entire operating systems for embedded computers as processes on a server. Contiki[13], or similar OSs, can be used to implement new software appliances from scratch. This approach cannot be used to interface existing programs (e.g., an existing web server like Apache) to a virtual network, unless the software interface for networking is completely rewritten to support virtual networking.

ViewOS[14] is a partial virtualization project. View-OS virtualizes the system calls generated by the programs, so unmodified binary programs can run in the virtualized environment. ViewOS supports the re-definition of the networking services at user level. Server, client and peer-to-peer programs can run transparently on a View-OS machine as if they were running just on the OS, but using a virtualized stack instead of the kernel stack.

Another project provides network virtualization in the NetBSD environment: Rump Anykernel[15]. The idea of Rump is to provide user mode environments where kernel drivers and services can run. Rump provides a very useful structure for kernel code implementation and debugging, as entire sections of the kernel can run unmodified at user level. In this way it is possible to test unstable code without the risk of kernel panic.

At the same time, Rump provides a way to run kernel services, like the TCP-IP stack, at user level. It is possible to reuse the kernel code of the stack as a networking library or as a networking deaemon at user level.

 ISBN: 978-1-61208-280-6

D. Multiple Stack support

Some IoTh applications require the ability for one process to be connected to several TCP-IP stacks at the same time. The Berkeley sockets API has been designed to support only a single implementation for each protocol family.

ViewOS and LWIPv6 use an extension of the Berkeley Socket API, msocket[16], providing the support for multiple protocol stacks for the same protocol family.

V. Usage Cases

This section describes some general usage cases of IoTh. A complete description of the experiments, including all the details to test the results, can be found in the Technical Report [17].

A. Client Side usage cases

- Co-existence of multiple networking environments. This feature can be used in many ways. For example, it is possible to have a secure VPN connected to the internal protected network of an institution or a company (an intranet) on which it is safe to send sensitive data and personal information, and a second networking environment to browse the Internet.
  As a second example, technicians who need to track networking problems may find it useful to have some processes connected to the faulty service, while a second networking environment can be used to look for information on the Internet, or to test the faulty network by trying to reach the malfunctioning link from the other end.
- Creation of networking environments for IPC. Many programs have web user interfaces for their configuration (e.g., CUPS or xmbc). Web interfaces are highly portable and do not require specific graphics libraries to run. Using IoTh it is possible to create several Local Host Networks (LHN), i.e., virtual networks for IPC only, to access the web interfaces of the running processes. LHN can have access protection, e.g., an LHN to access the configuration interface of critical system daemons can be accessible only by *root* owned processes. All daemons can have their own IP address, logical name and run their web based configuration interface using port 80.

B. Server side usage cases

- Virtual hosting is a well-known feature of several networking servers: the same server provides the same kind of service for multiple domains. IoTh generalizes this idea. It is possible to run several instances of the same networking daemon, giving each one its IP address. It is possible to run several pop, imap, DNS, web, MTA, etc.. deamons, each one using its own stack. All the deamons will use their standard port numbers.
- Service migration in IoTh is as simple as stopping the daemon process on one host and starting it on another one. In fact, a deamon process can have its embedded networking stack, so its IP address and its routing rules are just configuration parameters of the daemon process itself. A VDE can provide a virtual Ethernet for all the processes running on several hosts. Stopping the daemon process on one server and activating it later on a second server providing the same VDE is, in the virtual world, like unplugging the Ethernet cable of a computer from a switch and plugging it into a port of another switch of the same LAN (the ports of both switches have the same untagged VLAN).
- With IoTh it is possible to design network daemons which change their IP addresses in a dynamic way. One Time IP address (OTIP) applies to IP addresses the same technique used for passwords in One Time Password (OTP) services. In OTP, the password to access a service changes over time and the client must compute the current password to be used to access the service. This is common for protecting on-line operations on bank accounts. OTIP uses the same concept to protect private services accessible on the Internet. A deamon process changes its IP address dynamically over time and all its legitimate users can compute its current IP address using a specific tool, and connect. Port scan traces and network dumps cannot provide useful information for malicious attacks because all the addresses change rapidly. A public demo of OTIP was given during FOSDEM2012[18].

C. Other usage cases

IoTh allows us to use several networking stacks. These stacks can be several instances of the same stack, or different stacks. In fact, it is possible to have different implementations of TCP-IP stacks or stacks configured in different ways, available at the same time. Processes can choose which one is best suited to their activities.

This feature can be used in different ways:

- Using an experimental stack as the single, shared stack of a remote computer can partition the remote machine in cases of a malfunctioning of the stack itself. IoTh enables the coexistence of the stack under testing with a reliable production stack, which can be used as a safe communication channel.
- Processes can have different networking requirements. For example, communicating peers on a high latency link need larger buffers for the TCP sliding window protocol. It is possible to configure each stack and fine tune its parameters for the requirements of each process, as each process can have its own stack.

VI. Security Considerations

Several aspects of security must be taken into consideration in IoTh.

It is possible to limit the network access possibilities of an IoTh process and restrict the network services it can use. In fact, each IoTh process must be connected to a virtual local network to communicate, and virtual local networks have access control features. In VDE, for example, the permission to access a network is defined using the standard access control mechanisms of the file system. The interaction between

 ISBN: 978-1-61208-280-6

TABLE I
COMPARISON IN BANDWIDTH (MB/S) BETWEEN A KERNEL STACK AND IOTH

	10MB kernel	10MB IoTh	20MB kernel	20MB IoTh	40MB kernel	40MB IoTh
localhost	116	29.9	118	35.9	136	37.4
network 1Gb/s	104	41.9	112	49.0	112	51.7
network 100Mb/s	11.2	11.0	11.1	11.0	11.1	11.0

processes connected to a VDE and the other networks (or the entire Internet) can be regulated by specific configurations of the virtual routers used to interconnect that VDE.

It is also possible to consider the positive effects of IoTh with respect to protection from external attacks. Port scanning [19] is a method used by intruders to get information about a remote server, planned to be a target for an attack. A port scan can reveal which deamons are currently active on that server, then which security related bugs can be exploited.

This attack method is based on the assumption that all the deamons are sharing the same IP stack and the same IP addresses. This assumption is exactly the one negated by IoTh. Port scanning is almost useless in IoTh, since an IP address is daemon specific, so it would reveal nothing more than the standard ports used for that service. When IoTh is applied to IPv6, the process IP address on a VDE network can have a 64-bit prefix and 64 bits for the node address. A 64-bit address space is too large for a brute force address scan to be effective.

There are also other aspects of security to be considered regarding the effects of IoTh on the reliability of the hosting system. Daemon processes run as unprivileged user processes in IoTh. They do not even require specific capabilities to provide services on privileged ports (CAP_NET_BIND_SERVICE to bind a port number less than 1024). The less privileged a daemon process is, the smaller the damages it may cause in cases when the daemon is compromised (e.g., by a buffer overflow attack).

VII. PERFORMANCE OF IOTH

IoTh provides a new viewpoint on networking. As this paper has shown in the previous sections, IoTh allows a wide range of new applications. IoTh flexibility obviously costs in terms of performance. A fair analysis of IoTh performance has to consider the balance between the costs of using this new feature and the benefits it gives. In the same way processes run faster on an Operating System not supporting Virtual Memory, but, for many applications, the cost of Virtual Memory is worthwhile because you can run a greater number of processes. The IoTh approach can co-exist with the standard management of IP addresses and services. System administrators can decide which approach is more suitable for each service.

Table I shows the comparison of the bandwidth of a TCP connection between the Linux Kernel TCP-IP stack implementation and a IoTh implementation based on VDE and LWIPv6. The test set includes the measure of the bandwidth for file transfers of 10MB, 20MB and 40MB between processes running on the same host, on hosts connected by a 100Mb/s LAN and by a 1Gb/s LAN. The test environment consists of two GNU-Linux boxes (Debian SID distribution), Linux 3.2 kernel, NetXtreme BCM5752 controller, dual core Core2Duo processor running at 2Ghz, HP ProCurve Switches 1700 and 1810G. The files have been transferred using wget.

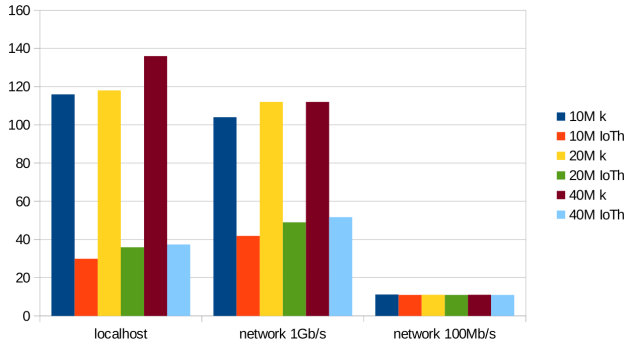


Fig. 2. A graphical view of Table I data

From the table and from the graph of Fig. 2 it is possible to see that IoTh can reach a sustained load of about 50MB/s, so the overhead added by the new approach is appreciable only on very fast communication lines. On a 100Mb/s LAN the difference is minimal. The improved performance for larger file transfers is caused by the constant startup cost (socket opening, http protocol, etc) which is distributed on a longer operation. On localhost or on fast networks, the bandwidth of IoTh is about a quarter to a half of the bandwidth reached by the kernel.

It is worth considering that, in this test, both VDE and LWIPv6 run at user level. These are the performance values of the less efficient implementation structure of IoTh. Kernel level implementations of the TCP-IP stack library, and of the virtual networking switch engine, will increase the performance of IoTh.

VIII. CONCLUSION AND FUTURE WORK

IoTh opens up a range of new perspectives and applications in the field of Internet Networking.

IoTh unifies the role of networking and IPC, so it can play an important role in the design of future applications: distributed applications and interoperating processes can use the same protocols to communicate.

The challenge of supporting new IoTh based services creates a need to analyze the TCP-IP protocols, in order to evaluate if and how these protocols, designed for physical networks, need to be modified or updated to be effective in IoTh. An example of a question that needs to be evaluated is whether the DNS protocol can have specific queries or features for IoTh.

On the other hand, IoTh requires an efficient infrastructure, able to provide a virtual networking (Ethernet) service to

 ISBN: 978-1-61208-280-6

processes. The research should also consider new efficient ways of interconnecting the local virtual networks to provide a better usage of virtual links, both for efficiency and for fault tolerance.

All the software presented in this paper has been released under free software licenses and has been included in the Virtual Square tutorial disk image [20]. This disk image can be used to boot a Debian SID GNU-Linux virtual machine. All the software tools and libraries used in this paper have already been installed and the source code of everything not included in the standard Debian distribution is also available in the disk image itself.

ACKNOWLEDGMENTS

I would like to express my gratitude to all the software designers and developers of the VirtualSquare Lab who have shared my daydreaming about the virtualization of everything, patiently following all my brainstorming.

REFERENCES

[1] J. Postel, "DoD standard Internet Protocol," RFC 760, Internet Engineering Task Force, Jan. 1980, obsoleted by RFC 791, updated by RFC 777. [Online]. Available: http://www.ietf.org/rfc/rfc760.txt 04.15.2013

[2] D. Price and A. Tucker, "Solaris zones: Operating system support for consolidating commercial workloads," in *Proceedings of the 18th USENIX conference on System administration*, ser. LISA '04. Berkeley, CA, USA: USENIX Association, 2004, pp. 241–254.

[3] LXC team, "lxc linux containers," http://lxc.sourceforge.net/ 04.15.2013.

[4] IEEE and The Open Group, "Posix.1 2008," http://pubs.opengroup.org/onlinepubs/9699919799/ 04.15.2013.

[5] K. Ashton, "That 'Internet of Things' thing," *RFID Journal*, vol. 22, pp. 97–114, 2009.

[6] R. Davoli, "Vde: Virtual distributed ethernet," in *Proceedings of the First International Conference on Testbeds and Research Infrastructures for the DEvelopment of NeTworks and COMmunities*, ser. TRIDENTCOM '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 213–220.

[7] Open vSwitch team, "Open vswitch," http://openvswitch.org/ 04.15.2013.

[8] L. Rizzo and G. Lettieri, "Vale, a switched ethernet for virtual machines," University of Pisa, Italy, Tech. Rep., 2012. [Online]. Available: http://info.iet.unipi.it/∼luigi/papers/20120608-vale.pdf 04.15.2013

[9] L. Rizzo, "Revisiting network i/o apis: the netmap framework," *Commun. ACM*, vol. 55, no. 3, pp. 45–51, 2012.

[10] A. Dunkels, "Full tcp/ip for 8-bit architectures," in *Proceedings of the 1st international conference on Mobile systems, applications and services*, ser. MobiSys '03. New York, NY, USA: ACM, 2003, pp. 85–98.

[11] A. Dunkels, L. Woestenberg, K. Mansley, and J. Monoses, "Lwip," http://savannah.nongnu.org/projects/lwip 04.15.2013.

[12] R. Davoli, "Lwipv6," http://wiki.virtualsquare.org/wiki/index.php/LWIPV6 04.15.2013, 2007.

[13] A. Dunkels, B. Gronvall, and T. Voigt, "Contiki - A Lightweight and Flexible Operating System for Tiny Networked Sensors," in *Proceedings of the 29th Annual IEEE International Conference on Local Computer Networks*, ser. LCN '04. Washington, DC, USA: IEEE Computer Society, pp. 455–462.

[14] L. Gardenghi, M. Goldweber, and R. Davoli, "View-os: A new unifying approach against the global view assumption," in *Proceedings of the 8th international conference on Computational Science, Part I*, ser. ICCS '08, 2008, pp. 287–296.

[15] A. Kantee, "Flexible operating system internals: The design and implementation of the anykernel and rump kernels," 2012, doctoral Dissertation, Aalto Univerisity, Finland.

[16] R. Davoli and M. Goldweber, "msocket: multiple stack support for the berkeley socket api," in *SAC '12: Proceedings of the 27th Annual ACM Symposium on Applied Computing*, 2012, pp. 588–593.

[17] R. Davoli, "Internet of threads, technical report," http://www.cs.unibo.it/∼renzo/iothtr.pdf 04.15.2013, Computer Science and Engineering Department. University of Bologna, Tech. Rep., 2013.

[18] ——, "Video of the public "internet of threads" demo, fosdem 2012," http://video.fosdem.org/2012/maintracks/k.1.105/Internet_of_Threads.webm 04.15.2013, 2012.

[19] Fyodor Vaskovich (Gordon Lyon), "The art of port scanning," *Phrack*, vol. 7, no. 51, 1997.

[20] R. Davoli, "Virtual square tutorial disk image," http://wiki.virtualsquare.org/wiki/index.php/Virtual_Square_Tutorial_Disk_Image 04.13.2013.

 ISBN: 978-1-61208-280-6

A Complexity Analysis of an XML Update Framework

Mohammed Al-Badawi and Abdallah Al-Hamadani
The Department of Computer Science
Sultan Qaboos University
Muscat, Oman
mbadawi@squ.edu.om, abd@squ.edu.om

***Abstract*—XML update is problematic for many XML database techniques. The main issue tackled by these techniques is the cost reduction of updating the XML's hierarchal structure inside the underlying storage. PACD technique, introduced earlier, is an attempt in this direction. This paper mainly provides a complexity analysis of the PACD's updates primitives. The analysis, along with the comparative experimental results presented here, have shown that the cost of eight update primitives (out of nine discussed) leys under acceptable range of a constant 'c' where 'c' is an extremely small number comparing to the number of nodes 'n' in the underlying database. Such good performance is lacked in the compared techniques.**

Keywords-XML Databases;XML Update; Mapping

I. INTRODUCTION

Data stored in the extensible markup language (XML) containers (database) is subject to updates when circumstances change. Unfortunately, handling XML updates is a common problem in the existing XML storage models and optimization techniques. Relational approaches using node labeling techniques [6][12][13][14][16][22] require a large number of renumbering operations in order to keep the node labels updated whenever a node is inserted, deleted or moved from one location to another in the XML tree. For those approaches which use path summaries to encode the XML hierarchical structure, e.g., [3][4][7], an additional cost results from updating these summaries. In native XML approaches such as sequence based [10][15][17] and feature based techniques [19][23], the update problem is even worse. In the first case, the consequences of a single update operation, for example deleting a node, can affect hundreds or even thousands locations in the corresponding sequence depending on the node's location in the XML tree. A similar problem occurs in the case of feature based techniques, which rely on encoding the relationship between the nodes and the different ePaths of the XML tree into what is called feature-based matrices [19].

PACD, an acronym for Parent-Ancestor-Child-Descendant, as an XML processing technique introduced in [2], brings the cost of updating the XML's hierarchal structure to the data representation level by encoding these structures into a set of structure-based matrices, which allow direct access to the information of the nodes affected by such update operations. This paper mainly introduces the PACD's Updates Query Handler (UQH) and provides a complexity analysis of its update primitives. The paper starts by revisiting the PACD's framework in Section II; then, it introduces the UQH framework in Section III. Section IV discusses the complexity of the different update primitives while Sections V and VI, respectively, summarize the complexity discussion and provides a supportive comparative experimental result. The paper is concluded in Section VII.

II. PACD'S XML PROCESSING MODEL

PACD, introduced in [1][2], is a bitmap XML processing technique consisting of two main components: the Index Builder (IB) and the Query Processor (QP). The IB (see Figure 1) shreds the XML's hierarchal structure (derived by the XPath's thirteen axes and their extension; the Next and Previous axes [1]) into a set of binary relations each of which is physically stored as an n×n bitmap matrix. An entry in any matrix is either '1' if the corresponding relationship is exists between the coupled nodes or '0' otherwise [8][23].

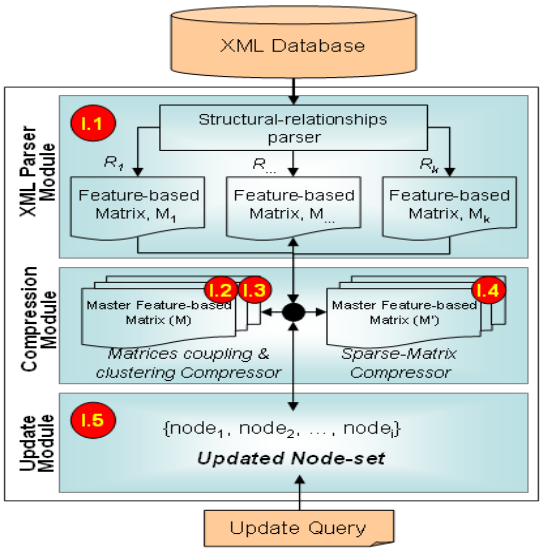


Figure 1. PACD's Index Builder (IB)

The IB component is also responsible to handle the XML updates. Once an update query is issued, the attached UQH determines the nodes affected by the update query and the type of the update operation itself. The following section introduces the PACD's UQH and its update primitives, while the primitives' complexity is discussed in Section IV.

 ISBN: 978-1-61208-280-6

III. THE UPDATE HANDLER

PACD's Update Query Handler (UQH), represented by operation I.5 of Figure 1, is responsible for all update tasks targeted by any ML update operation. These tasks include the translation of the update query, and the determination and the execution of necessary update primitive.

The execution of the update query starts by identifying the node(s) that are affected by the update command/query. The handler navigates through a finite-state-machine (FSM) version of the update-query [2] in order to locate the affected node-set. Once the target node-set is identified, the handler calls the appropriate update primitive (see Table I). PACD supports update primitives for single node insertion and deletion, twig insertion and deletion, and textual and structural-based changes.

The update primitives act on all PACD's components including the NodeSet container [2] and the structure based matrices. Each update primitive has to execute certain instructions over each component such as adding new columns and rows (over the bitmapped matrices). The cost of the update-query is the sum-cost of executing all generated update primitives over the all PACD's components, i.e., matrices and NodeSet container. For example, the 'insert' primitive can involve adding one or more rows and columns to the bitmapped matrices, as well as adding one or more entries to the NodeSet container. So, the cost of the insert operation will be the cost of inserting the node information inside the NodeSet container plus the cost of inserting one row and column inside the 'child', 'desc' and 'next' matrices respectively. The general update algorithm is given in Table II.

TABLE I. XML UPDATE PRIMITIVES

Insertion	insertLeaf	adds a leaf node
	insertNonLeaf	adds an internal node
	insertTwig	adds a single-rooted, connected sub-tree
Deletion	deleteLeaf	removes a leaf node
	deleteTwig	remove a single-rooted, connected sub-tree
Updating	changeName	rename an element or attribute name
	changeValue	edit the value (text) of an attribute (element)
	shiftNode	move a node from one place to another
	shiftTwig	move a single-rooted, connect sub-tree from one place to another

TABLE II. XML UPDATES EXECUTION ALGORITHM

```
INPUT: update-query
OUTPUT: none
Construct the FSM execution plan corresponding to query's twig
node-set = the returned node-set from the FSM execution
Using the update-query syntax, determine the primitive(s)
Call the update-primitive(s) with the obtained node-set:
    Alter the NodeSet container;
    Alter the childOf matrix;
    Alter the descOf matrix;
    Alter nextOf matrix;
End;
```

IV. DISCUSSION OF UPDATE PRIMITIVES

This section discusses, through examples, the complexity of the update primitives. The complexity is counted as the number of the 'change' actions performed on the NodeSet, child matrix, descendant (desc) matrix and next matrix. Due to space limitation, the paper only presents sample algorithms for some update primitives and then provides an example on how the algorithm works based on the XML tree in Figure 2.

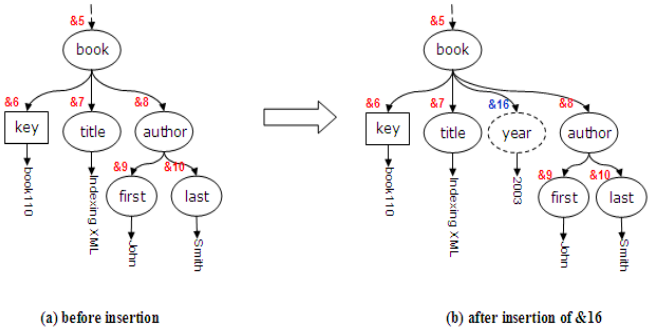


Figure 2. An Example XML Tree (includes an insertion case)

A. *Insert Primitives*

Table I has shown three insertion primitives that can be triggered against XML databases. The prototypes of the methods implementing these primitives are:

insertLeaf(node_info, parentID [,precedingID]);
insertNonLeaf(node_info, parentID [,precedingID]);
insertTwig(twig_info, parentID [,precedingID])

In the above prototypes, the 'node_info' indicates the information of the node(s) to be inserted which includes the nodeID, tag_name and the optional textual contents. The 'parentID' refers to the node ID of the parent node under which the insertion will take place. The 'precedingID' must be specified if the document order [6][16] is to be preserved, and it indicates the node ID of the node that must precede the new node.

1. *Leaf Node Insertion*

This method inserts a node at the bottom-most level of the tree under parentID node and next to precedingID node. Both parentID and precedingID are identified by the UQH during the query translation process.

Example: Add the 'year' information, i.e., 2003, to the book identified by the key 'book/110', where the 'year' information must precede the 'author' information (Figure 2).

The cost breakdown of the above operation is:

NodeSet	child	desc	next	Total
1	3	4	4	12 hits

2. *Non-Leaf Node Insertion*

This method can insert a node at any level of the tree except the bottom-most level. ParentID and precedingID are

identified by the UQH prior calling the primitive. In this paper, the analysis assumes that the primitive is only creating an additional level between a parent and its children, making these children as the children of the inserted node.

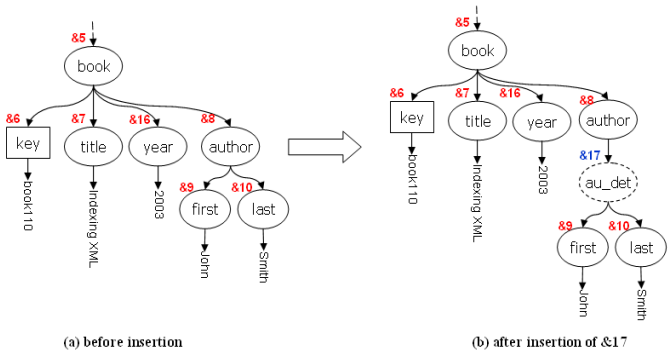
TABLE III. INSERTING NON-LEAF NODE ALGORITHM

```
insertNonLeaf(node_info:nodeType,parentID:nodeIDType,precID:
nodeIDType)
   Get the next nodeID;
   Insert the node information into NodeSet;
   *-- update the child matrix:
   Add a row and column to the 'child';
   Let: childSet = {node(i), where child[i,parentID] = '1'}
   For each i ∈ childSet:
       Set: child[i,nodeID] = '1';
   Set: child[nodeID,parentID] = '1';
   *--update the desc matrix:
   Add a row and column to the 'desc';
   Let: anceSet = {node(i), where desc[parentID,i] = '1'} ∪
       parentID;
   Let: descSet = {node(j), where desc[j,parentID] = '1'};
   For each i ∈ anceSet:
       Set: desc[nodeID,i] = '1';
   For each j ∈ descSet:
       Set: desc[j,nodeID] = '1';
   *--update the next matrix:
   Add a row and column to the 'next';
   If precID ≠ null:
       Let: temp = {node(i), where next[i,precID] = '1'};
       Set: next[nodeID,precID] = '1';
       If temp ≠ null:
           Set: next[temp,precID] = '1';
END.
```

Example: Make the current author of the book titled 'Indexing XML' to be the FIRST author of the book so that other authors can be added. This requires adding a parent node called 'au_det' for the 'first' and 'last' nodes under the original 'author' node (Figure 3).

The cost breakdown of the above operation is:

NodeSet	child	desc	next	Total
1	5	6	0	12 hits



(a) before insertion (b) after insertion of &17

Figure 3. Non-Leaf Node Insertion

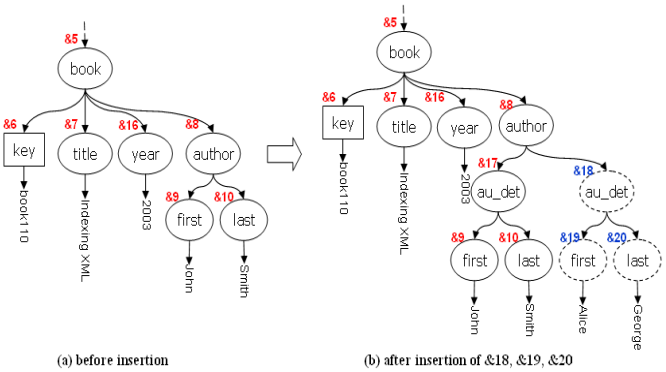
3. *Twig Insertion*

This method inserts a sub-tree of 'm' nodes under the parentID and after the precID. Both the parentID and the precID are determined by the UQH, and the twig is only inserted at bottom-most nodes. The twig insertion can be modeled as inserting multiple connected nodes. In other words, inserting a twig of 'm' nodes requires 'm' times the cost of inserting a single leaf-node and can be performed by the same algorithm in Table III starting at the twig's root node.

Example: add a second author sub-tree, i.e., including the 'first' and 'last' name, to the book titled 'Indexing XML' (Figure 4).

The cost breakdown of the above operation is:

NodeSet	child	desc	next	Total
3	9	14	8	34 hits



(a) before insertion (b) after insertion of &18, &19, &20

Figure 4. Twig Insertion

4. *Complexity Analyses*

Due the space limitation the full complexity analyses of the operations is omitted. The following table only summarizes the number of work-units required to conduct the insertion primitives in general.

TABLE IV. COMPLEXITY ANALYSES SUMMARY OF THE INSERTION PRIMITIVES

Operation	Growth in		# of Work-Units (Hits)				
	NodeSet	Matrix	NodeSet	childOf	descOf	nextOf	Max. Complexity
insertLeaf	1 rec. more	1 row more 1 col more	1	2+1	2+h	2+2	O(c)
insertNonLeaf	1 rec. more	1 row more 1 col more	1	2+α	2+f×n	2+2	O(f×n)
insertTwig (m nodes)	m rec. more	m row more m col more	m	m.(2+1)	m.(2+h)	m.(2+2)	O(m.c)

n= total number of nodes in the XML tree
h= the maximum height of the XML tree (# of levels)
α= the maximum breadth-degree (i.e. number of children) of any XML node
f= a number between 0 and 1, where 'f×n' is the number of descendants at any node, usually $0 \le f \le ½$
c= is very small number comparing to 'n' in large XML databases such that $\lim_{n\to\infty} \frac{c}{n} = 0$

B. *Deletion Primitives*

Table I has shown two deletion primitives that can be triggered against XML databases. The prototypes of the methods implementing these primitives are:

deleteLeaf(nodeID);

deleteTwig(twigRootNodeID);

In the above prototypes, the 'nodeID' indicates the node ID of the node that to be deleted while the 'twigRootNodeID' indicates the node ID of the root node of the targeted twig. The discussion in this section assumes that deleting a non-leaf node results in a cascade deletion of its children, therefore the 'deleteTwig' operation will be applied in the case of deleting a non-leaf node.

1. *Leaf Node Deletion*

This method deletes a node from the bottom-most level of the tree labeled with nodeID, which is returned by the UQH during the query translation process.

Example: Remove the author's last-name from the book identified by the key 'book/110'.

The cost breakdown of the above operation is:

child	desc	next	NodeSet	Total
2	2	2	1	7 hits

2. *Twig Deletion*

This method deletes a connected sub-tree rooted at 'twigRootNodeID' from the XML tree. The twig root node ID is returned by the UQH during the query translation process.

TABLE V. INSERTING NON-LEAF NODE ALGORITHM

```
deleteTwig(twigRootNodeID: nodeIDType)
  *-- reconnect the next_of list of the nextOf matrix:
  Let:
      next = {node(i), where nextOf[i,twigRootNodeID] = '1'};
      prev = {node(j), where nextOf[twigRootNodeID,j] = '1'};
  If next ≠ null AND prev ≠ null:
      Set: nextOf[next,prev] = '1';
  *--identify all the node inside the deleted twig:
  Let: descSet = {node(i), where descOf[i, twigRootNodeID] =
     '1'} ∪ twigRootNodeID;
  *--remove row and columns from all matrices, and the node_info
  *-- from the NodeSet :
  For each i ∈ descSet:
      Locates the corresponding row and column of the nodeID
          inside the 'childOf';
      Remove the row and column from the 'childOf';
      Locates the corresponding row and column of the nodeID
          inside the 'descOf';
      Remove the row and column from the 'descOf';
      Locates the corresponding row and column of the nodeID
          inside the 'nextOf';
      Remove the row and column from the 'nextOf';
      Locate the corresponding record of the nodeID inside the
          'NodeSet';
      Delete the nodeID;
END.
```

Example: Remove the complete author's information from the book identified by the key 'book/110' (Figure 5). Note: this will remove the nodes '&8' and '&9'.

The cost breakdown of the above operation is:

child	desc	next	NodeSet	Total
4	4	4	2	14 hits

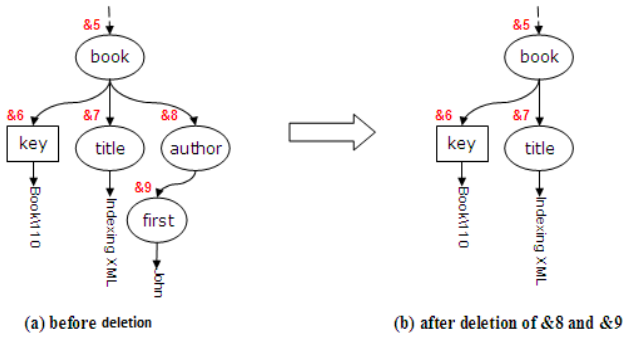


Figure 5. Twig Deletion

3. *Complexity Analyses*

The following table only summarizes the number of work-units required to conduct the deletion primitives in general.

TABLE VI. COMPLEXITY ANALYSES SUMMARY OF THE DELETION PRIMITIVES

Operation	Growth in		# of Work-Units (Hits)				
	NodeSet	Matrix	NodeSet	childOf	descOf	nextOf	Max. Complexity
deleteLeaf	1 rec. less	1 row less 1 col less	1	2	2	2+1	O(c)
deleteTwig (m nodes)	m rec. less	m rows less m cols less	m	m×2	m×2	m×(2+1)	O(m.c)
n= total number of nodes in the XML tree c= is very small number comparing to 'n' in large XML databases such that $\lim_{n\to\infty} \frac{c}{n} = 0$							

C. *Change Primitives*

Table I has shown four change primitives that can be triggered against XML databases. The prototypes of the methods implementing these primitives are:

changeName([nodeID|oldName],newName);
changeValue([nodeID|oldName],newValue);
shiftNode(nodeID,newParentID[,leftID]);
shiftTwig(twigRootID,newParentID[,leftID]);

In the above prototypes, the 'nodeID' indicates the node ID of the targeted node targeted. The 'oldName' and 'newName' indicate the tag-name of the targeted node. The 'newValue' is the textual content of the node to be altered. The 'parentID' and the 'leftID' are the node ID of the parent node and left node of the targeted node. Finally, the 'twigRootID' is the node ID of the twig to be shifted.

1. *Tag-Name Change*

This method renames a node (identified by the nodeID) or a set of nodes (that have the same name identified by oldName) to the new name newName.

Example: Change the name of the node 'thesis' to be 'phdthesis'.

This query changes the tag name of the node &11 from 'thesis' to 'phdthesis' with the cost of one work-unit.

Example: Change the name of all nodes labeled with 'key' to be 'pub_id'.

 ISBN: 978-1-61208-280-6

In this query, the 'oldName' parameter is the word 'title' and the 'newValue' parameter is a function that converts its argument to the uppercase. The query will perform three work units in total.

2. *Textual-Value Change*

This method changes the textual contents of a node (identified by the nodeID) or a set of nodes (that have the same name identified by oldName) to the new value newValue.

Example: Change the publication year for the book labeled with 'Book/101' to be '2000' instead of '2001'.

This query changes the value of the node &2 from '2001' to '2000' with the cost of one work-unit.

Example: Change the 'title' of all publications to the uppercase.

In this query, the 'oldName' parameter is 'title' and the 'newValue' parameter is a function that converts its argument to the uppercase. The query will perform three work units in total.

3. *Single Node Shifting*

This method moves the node labeled with nodeID to be under the node newParentID. If the exact location is required, the preceding node at the new location, i.e., 'leftID', must be specified.

Example: Move the publication year of book 'book/101' to be the publication year for the book 'Book/110' (Figure 6).

The cost breakdown of the above operation is:

childOf	descOf	nextOf	Total
2	4	4	10 hits

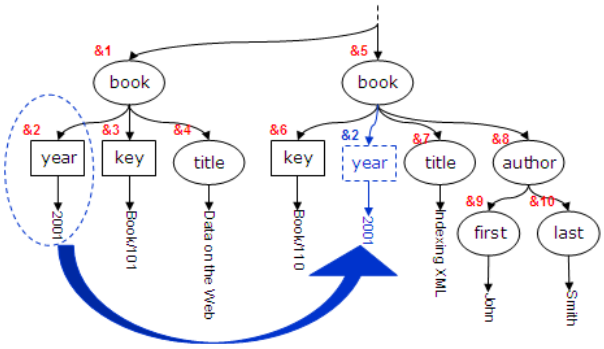


Figure 6. Single Node Shifting (Parent Change)

4. *Twig Shifting*

This method moves a sub-tree (twig) rooted at the twigRootID to be a sub-tree under the node newParentID. If the exact location is required, the preceding node at the new location, i.e., leftID, must be specified.

TABLE VII. INSERTING NON-LEAF NODE ALGORITHM

```
shiftTwig(twigRootID: nodeIDType, newParentID: nodeIDType,
                leftID: nodeIDType)
  *-- update the childOf matrix:
  Let: oldParentID = {node(i), where childOf[twigRootID,i] =
       '1'};
  Set:
      childOf[twigRootID,newParentID] = '1';
      childOf[twigRootID,oldParentID] = '0';
  *--update the descOf matrix:
  Let:
      twigNodeSet = {node(1..m), where node(i) ∈ twig};
      oldAnceSet = {node(i), where descOf[twigRootID,i] = '1'};
      newAnceSet = {node(j), where descOf[newParentID,j] = '1'}
          ∪ newParentID;
  For each node i ∈ newAnceSet:
      For each node j ∈ twigNodeSet:
          Set: descOf[j,i] = '1';
  For each node i ∈ oldAnceSet:
      For each node j ∈ twigNodeSet:
          Set: descOf[j,i] = '0';
  *--update the nextOf matrix:
  Let:
      next_of_ twigRootID = {node(i), where
          nextOf[i, twigRootID] = '1'};
      prev_of_ twigRootID = {node(j), where
          nextOf[twigRootID,j] = '1'};
      next_of_leftID = {node(i), where nextOf[i,leftID] = '1'};
      prev_of_leftID = {node(j), where nextOf[leftID,j] = '1'};
  Set (if any combination is not null):
      nextOf[next_of_twigRootID,prev_of_twigRootID] = '1';
      nextOf[twigRootID,prev_of_twigRootID] = '0';
      nextOf[twigRootID,leftID] = '1';
      nextOf[next_of_leftID, twigRootID] = '1';
      nextOf[leftID,prev_of_leftID] = '0';
      nextOf[next_of_leftID,leftID] = '0';
END.
```

Example: Move the author information of book 'book/110' to be the author for the book 'Book/101' (Figure 7).

The cost breakdown of the above operation is:

child	desc	next	Total
2	12	2	16 hits

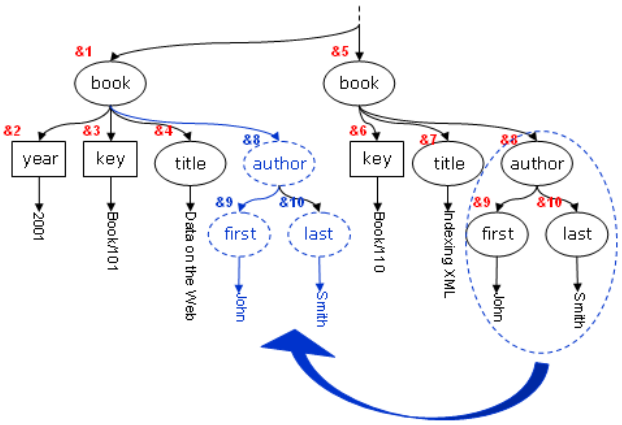


Figure 7. Twig Shifting (Parent Change)

5. *Complexity Analyses*

The following table only summarizes the number of work-units required to conduct the change primitives in general.

TABLE VIII. COMPLEXITY ANALYSES SUMMARY OF THE CHANGE PRIMITIVES

Operation	Growth in		# of Work-Units (Hits)				
	NodeSet	Matrix	NodeSet	childOf	descOf	nextOf	Max. Complexity
chnageName	none	none	1 or k	0	0	0	O(k)
changeValue	none	none	1 or k	0	0	0	O(k)
nodeShift	none	none	0	2	2×h	6	O(c+2.h)
twigShift (m nodes)	none	none	0	m×2	m×2×h	6	O(c+m×2×[h+1])

n= total number of nodes in the XML tree
k= the number of nodes per tag/attribute name (usually much smaller than 'n')
h= the height of the XML tree (# of levels)
c= is very small number comparing to 'n' in large XML databases such that $\lim_{n\to\infty} \frac{c}{n} = 0$

V. OVERALL COMPLEXITY

The analysis provided above has shown that the cost of all update-primitives over the PACD's *uncompressed* data representation lies in acceptable limits in general. Of the update primitives discussed, the highest update complexity is only a fraction of the number of nodes, i.e., 'n', and this only happens during the infrequently accessed operation 'insertNonLeaf'. The cost of other update operations ranges between a very small constant 'c' (where 'c' is an extremely small number comparing to the number of nodes 'n' in the database) and 'm×c' in the case of manipulating a *twig* of size 'm' nodes.

From the technical point of view, the bitmapped XML structure and the introduction of the previous/next axes [2] has played a major role in such cost reduction. Unlike node-labeling based techniques, e.g., [11][18][21], the use of the next matrix -to encode the document order- has narrowed the spread of label changes to the adjacent nodes (only) of the targeted node. Also encoding the basic XML structures (the child/parent and descendant/ancestor relationships) using the bitmapped node couplings (the child and desc matrices) has reduced the high cost and complexity that result from using path-summaries [5][9] [7][20] and sequences [10][15] to encode such structures. The analysis has shown that the number of changes in the child structure is bounded by a small constant 'c' (where 'c' is an extremely small number comparing to the number of nodes 'n' in the database) in most cases except the 'insertNonLeaf' primitive which requires 'α' number of hits depending on the node's breadth degree.

Another source of cost reduction in PACD's update transactions is the separation between the textual contents representation and the XML hierarchal structure representation. Because of that, the content-based primitives only affect the NodeSet container while the structure-based update primitives affect the bitmapped matrices. This is not the case of path-summary and sequence-based techniques where the underlying path-summary or sequence has to be changed for either type of updates (content-based or structure-based). In general, the number of hits over the NodeSet container is limited by the number of targeted nodes except when amending a tag/attribute name or a node value for a set of nodes that share the same tag/attribute name. In this case, the cost is limited by the number of nodes that share the same tag/attribute name which is also considered small comparing to the entire XML tree.

VI. A COMPARATIVE STUDY

A. *The Experiment Setup*

A comparative experiment between the performance PACD technique and two representative XML techniques from the literature is conducted to support the above complexity analyses. The experiment executes 6 update queries –as a representation of the above update primitives- translated over 3 XML databases for the 3 selected XML techniques. The 6 update queries are listed in Table IX while the characteristics of the 3 XML databases are given Table X. Also Table XI shows the XML/RDBMS mapping schema of the three compared techniques, PACD, XParent and Edge, while other specifications of these techniques can be found at the references [2], [9] and [24] respectively.

The experiment counts the number of changes (hits) done over the technique's underlying representation (2nd column of Table XII) of the XML database, and lists them per query ID in separate columns (see Table XII) over each XML database. The number of hits, over all components, is summed up in the last 3 rows of Table XII which summarizes the overall experimental results.

TABLE IX. THE EXPERIMENTAL UPDATE QUERIES

Query ID	Query Description
U1	Insert an Atomic Value, i.e., leave node
U2	Insert a Non-atomic Value, i.e., non-leave or internal node
U3	Delete an Atomic Value , i.e., leave node
U4	Delete a Non-atomic Value , i.e., non-leave or internal node
U5	Change an Atomic Value, i.e., the textual content of a node
U6	Change a Non-atomic Value, i.e., tag-name

TABLE X. FEATURES OF THE EXPERIMENTAL XML DATABASES

	DBLP [25]	XMark [27]	Treebank [26]
Size (#of nodes)††	2,439,294	2,437,669	2,437,667
Depth(#of levels)	6	10	36
Min Breadth†	2	2	2
Max Breadth	222,381	34,041	56,385
Avg Breadth†	11	6	3
#of Elements	2,176,587	1,927,185	2,437,666
#of Attributes	262,707	510,484	1

† Figures exclude leaf nodes

†† Dataset also contains two versions of each database at 50% and 25% of the size of the base database. Both the depth and the average breadth of the base databases are maintained in the smaller databases

 ISBN: 978-1-61208-280-6

TABLE XII. THE EXPERIMENTAL RESULTS

Tech. Name	Query / Tables	DBLP						XMark						Treebank					
		U1	U2	U3	U4	U5	U6	U1	U2	U3	U4	U5	U6	U1	U2	U3	U4	U5	U6
Edge	edge	6	1	5	78815	1	213634	2	1	1	792	1	80316	2	1	1	9	1	136545
PACD	XMLSym	0	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	0	1
	XMLNodes	1	1	1	13	0	0	1	1	1	48	0	0	1	1	1	8	0	0
	XMLValues	1	0	1	12	1	0	1	0	1	24	1	0	1	0	1	4	1	0
	OIMatrix	2	n	2	26	0	0	4	n	4	514	0	0	9	n	9	186	0	0
	nextOf	2	0	2	11	0	0	1	0	2	2	0	0	2	0	2	5	0	0
XParent	elem	6	1	4	78815	0	0	2	1	1	48	0	0	2	1	2	9	0	0
	data	6	0	4	12	1	0	2	0	1	24	1	0	2	0	2	4	1	0
	labelpath	0	146	0	0	0	8	0	252	0	0	0	9	0	338750	0	0	0	248480
	dataPath	1	1	1	13	0	0	1	1	1	48	0	0	1	1	1	9	0	0
	ancestor	2	n	2	26	0	0	4	n	4	514	0	0	9	n	9	186	0	0
Edge	Total	6	1	5	78815	1	213634	2	1	1	792	1	80316	2	1	1	9	1	136545
PACD	Total	6	n+2	6	62	1	1	7	n+2	8	588	1	1	13	n+2	13	203	1	1
XParent	Total	15	n+148	11	78866	1	8	9	n+254	7	634	1	9	14	n+338752	14	208	1	248480

n=2437669, is the number of nodes inside each XML database and it should be unified for all databases

B. *Experimental Results Summary*

Comparing to other techniques, PACD appeared having the best performance for most of the queries in all situations. The experiment has also shown that the performance of XParent and Edge was delayed by the cost of the document order persevering mechanism. PACD eliminates this cost by encoding the previous/next relationship which requires at most two amendments for any node update operation.

TABLE XI. THE EXPERIMENTAL COMPARABLE XML TECHNIQUES

Technique	Components (XML/RDBMS Mapping Schema)
PACD	XMLNodes(nodeID, type, tagID) XMLSym(tagID, desc) XMLValues(nodeID, value) childOf(childID, parentID) OIMatrix(Source, Target, relType) descOf(descID, anceID) nextOf(nextID, prevID)
Edge	Edge(source, target, ordinal, label, flag, value)
XParent	labelPath(pathID, length ,PathDesc) element(pathID, ordinal, nodeID) data(pathID, ordinal, nodeID, value) dataPath(nodeID, parented) ancestors(nodeID, anceID, level)

VII. CONCLUSION

This paper has discussed the PACD's updating framework which is managed by a set of low cost update primitives. Once an update query is issued, the Update Query Handler (UQH) process identifies the target node-set and the necessary update primitive(s). The translation of an update query may generate one or more update primitives each of which may alter one or more XML nodes. The UQH currently can generate nine update primitives divided into three categories; the insert, delete, and change primitives.

The analysis and the experimental results in this paper have shown that the computation cost of XML update queries can be improved using the update primitives, which specifically act on the PACD data representation. The cost analysis of all update primitives is provided in Tables IV, VI and VIII.

REFERENCES

[1] M. Al-Badawi, H. Ramadhan, S. North, and B. Eaglestone, "A performance evaluation of a new bitmap-based XML processing approach over RDBMS", Int. J. of Web Engineering and Technology, vol. 7, no. 2 , 2012, pp. 143 – 172.

[2] M. Al-Badawi, B. Eaglestone, and S. North, "PACD: A Bitmap-based Approach for Processing XML Data", WebIST'09, Lisbon, Portugal, 2009, pp. 66-71.

[3] Q. Chen, A. Lim, and K. Ong, "D(K)-Index: An adaptive structural summary for graph-structured data', In proceedings of the 2003 ACM SIGMOD international conference on Management of data, CA, USA, 2003, pp. 134-144.

[4] C. Chung, J. Min, and K. Shim, "APEX: An adaptive path index for XML data", In proceedings of the 2002 ACM SIGMOD international conference on Management of data, Madison, Wisconsin, 2002, pp. 121-132.

[5] R. Goldman, and J. Widom, "DataGuides: Enabling query formulation and optimaization in semistructured database", In proceedings of the 23rd international conference on VLDB, 1997, pp. 436-445.

[6] T. Härder, M. Haustein, C. Mathis, and M. Wagner, "Node labelling schemes for dynamic XML documents reconsidered" International Journal of Data Knowledge Engineering, vol. 60, I. 1, 2007, pp. 126-149.

[7] S. Haw, and C. Lee, "Extending path summary and region encoding for efficient structural query processing in native XML databases", Journal of Systems and Software, vol. 82, I. 6, 2009, pp. 1025-1035.

[8] H. He, H. Wang, J. Yang, and P. Yu, "Compact reachability labeling for graph-structured data", In proceedings of the 14th ACM international conference on Information and knowledge management, Bremen, Germany, 2005, pp. 594-601.

 ISBN: 978-1-61208-280-6

[9] H. Jiang, H. Lu, W. Wang, and J. Yu, "XParent: An efficient RDBMS-based XML database system", International conference on Data Engineering, CA, USA, 2002, p. 2.

[10] J. Kwon, P. Rao, B. Moon, and S. Lee, "Fast XML document filtering by sequencing twig patterns", ACM Transactions on Internet Technology (TOIT), vol. 9, I. 4, Article 13, 2009, pp. 13.1-13.51.

[11] J. Lu, T. Ling, C. Chan, and T. Chen, "From region encoding to extended dewey: On efficient processing of XML twig pattern matching", In proceedings of the 31st International Conference on VLDB, Trondheim, Norway, 2005, pp. 193-204.

[12] P. O'Neil, E. O'Neil, S. Pal, I. Cseri, G. Schaller, and N. Westbury, "ORD-PATHs: Insert-friendly XML node labels", In proceeding of ACM/SIGMOD international conference on Management of Data, 2004, pp. 903-908.

[13] W. Shui, F. Lam, D. Fisher, and R. Wong, (2005) "Querying and marinating ordered XML data using relational databases", Proceedings of the 16th Australasian database conference - vol. 39, Newcastle, Australia, 2005, pp. 85-94.

[14] I. Tatarrinov, S. Viglas, K. Beyer, J. Shanmugasundaram, E. Shekita, and C. Zhang, "Storing and querying ordered XML using a relational database system", ACM/SIGMOD Record, Madison, Wisconsin, 2002, pp. 204-215.

[15] H. Wang, and X. Meng, "On sequencing of tree structures for XML indexing", In the proceedings of the 21st international conference on Data Engineering, 2005, pp. 372-383.

[16] H. Wang, H. He, J. Yang, P. Yu, and J. Yu, "Dual labeling: Answering graph reachability queries in constant time", In the proceedings of the International conference of Data Engineering, 2006, pp. 75-86.

[17] H. Wang, X. Wang, and W. Zeng, "A research on automaticity optimization of KeyX index in native XML database", In proceedings of the 2008 international conference on Computer Science and Software Engineering, 2008, pp. 700-703.

[18] X. Wu, M. Lee, and W. Hsu, "A prime number labeling scheme for dynamic ordered XML trees", In proceedings of the 20th international conference on Data Engineering, 2004, pp. 66-78.

[19] J. Yoon, S. Kim, G. Kim, and V. Chakilam, "Bitmap-based indexing for multi-dimensional multimedia XML document", In proceedings of the 5th International Conference on Asian Digital Libraries-ICADL2002, Singapore, 2002, pp. 165-176.

[20] M. Yoshikawa, T. Amagasa, T. Shimura, and S. Uemura, "XRel: A path-based approach to storage and retrieval of XML documents using relational databases", ACM/IT., vol. 1, I. 1, NY, USA, 2001, pp. 110-141.

[21] J. Yun, and C. Chung, "Dynamic interval-based labelling scheme for efficient XML query and update processing", Journal of Systems and Software, vol. 81, I. 1, 2008, pp. 56-70.

[22] C. Zhang, J. Nsughton, D. DeWitt, Q. Luo, and G. Lohman, "On supporting containment queries in relational database management systems", In proceedings of the 2001 ACM SIGMOD international conference on Management of Data, California, USA, 2001, pp. 425-436.

[23] N. Zhang, M. Özsu, I. Ilyas, and A. Aboulnaga, "FIX: Feature-based indexing technique for XML documents", In proceedings of the 22nd international conference on VLDB, vol. 32, Seoul, Korea, 2006, pp. 259-270.

[24] D. Florescu, and D. Kossmann "A Performance Evaluation of alternative Mapping Schemas for Storing XML Data in a Relational Database", TR:3680, May 1999, INRIA, Rocquencourt, France, pp. 1-24.

[25] DBLP. The DBLP Website, Available at http://dblp.uni-trier.de/, [Last accessed on 28/04/2013].

[26] PennProj. The Penn Treebank Project Website, Available online at http://www.cis.upenn.edu/~treebank/, [Last accessed on: 28/04/2013].

[27] A. Schmidt, F. Waas, M. Kersten, D. Carey, I. Manolescu, and R. Busse. "XMark: A Benchmark for XML Data Management", International conference on Very Large Data Bases, Hong Kong, China, 2002, pp. 974-985.

 ISBN: 978-1-61208-280-6

Portable Full-text Retrieval System

Takehiko Murakawa Tatsuya Takehara
Faculty of Systems Engineering
Wakayama University
Wakayama, Japan
e-mail: takehiko@sys.wakayama-u.ac.jp

***Abstract*—We developed a portable full-text search system that runs on Microsoft Windows. The server programs, full-text search executables, and the PHP system files can be stored on a removable memory device, such as a USB flash drive. The storage location of documents and the index are assignable to the host PC or the USB drive. A browser-based crossover search can be performed on plain text files, Word and PDF documents, and Excel spreadsheets. While the system employs established executable files for the server programs and the full-text search engine, we developed interfaces to support system use. We tested the system with more than tens of thousands of records to verify immediate retrieval.**

Keywords-Web application development; full-text search; document retrieval; mobility

I. INTRODUCTION

The most desirable environment for document retrieval is one that allows quick and easy file retrieval. In addition, some users require the privacy, i.e., the protection of files, and some desire portable document retrieval to use in internet-incapable locations.

Personal full-text search cannot be realized exclusively by advancing internet technologies. For example, the field of humanities, many researchers find internet resources insufficient. However, they do retain text data in spreadsheets or document files to read later. These researchers genuinely require simple and functional full-text search and retrieval.

To satisfy mobility requirements, we have been developing a portable full-text retrieval service. In this paper, we describe a file retrieval system that is stored on a universal serial bus (USB) flash drive.

The system can function on a low-performance personal computer (PC) such as a netbook or laptop, running Windows XP or higher operating system (OS). The portable edition of XAMPP server programs and Hyper Estraier search engine reside on the USB flash drive. The registered documents are indexed by n-grams; therefore the search is language independent. Even though a file may contain multiple languages, the relevant documents can be located if an appropriate search term is used. Search results are presented as hyperlinks to documents that include the keyword. The system allows crossover searches of plain text files, Microsoft Word documents, Excel spreadsheets, and Portable Document Format (PDF) files.

The proposed system is an enhanced version of previously developed system [1]. While the fundamental features, such as file registration and retrieval, are consistent with those of the previous system, the proposed system employs portable executable files, including an open source Hypertext Transfer Protocol (HTTP) server, and MySQL, a database management system (DBMS). With careful selection of hardware and software and refinement of existing user interfaces, the service can be easily used on familiar Windows PCs.

The remainder of this paper is organized as follows. Section II explores the background of the full-text retrieval service. Section III details the services and software related to our system. Section IV describes the main body of the system, including hardware and software configuration, and provides details about several key features. Section V presents an evaluation of the system. Section VI acknowledges challenges and provides a general discussion to identify the significance of the system. Finally, Section VII presents conclusions and future work.

II. PRELIMINARIES

This section introduces the execution environment and defines key components of full-text search.

A. Execution Environment

We assume that humanities research consists of fieldwork and laboratory investigations. Researchers often visit temples, museums, or archive warehouses to examine unpublished historical material. Since some of the material might be unidentified, the researchers examine it with reference to known information. We have been developing support systems, primarily Web applications, for humanities researchers [1] [2] [3]. Currently, our system is aimed at full-text search of data and spot notations for subsequent full-text searches.

The data stored and searched in our system have some common characteristics, and researchers have common expectations. The content to be searched is usually in plain text format. Although researchers do encounter data in rich document formats, such as Word or PDF, they do not expect mechanical retrieval of figures or complicated structures. Since the researchers are interested in historical material that describes the human environments and events that occurred approximately 1,000 years ago in Japan, the documents are primarily written using Chinese characters. In addition, the users may wish to add a notation in Japanese to the material they locate and will expect to locate this material by searching for the notation. Our system targets researchers who are adept at using PCs and who understand what sort of

 ISBN: 978-1-61208-280-6

data is suitable for computer processing. The researchers' basic tasks are (1) to read the freshly discovered documents, (2) to add one or more notations to the document to be used by our retrieval service, (3) to find and read relevant documents, and (4) to estimate the age and/or verify the validity of the document.

B. Full-text Search

In this section, we define several common terms specific to the full-text search described in this paper. We define a *search engine* as software that performs a full-text search. The process that involves a search engine and search interfaces is called a *retrieval service*. Google and Yahoo! are global retrieval services for internet resources. Note that we are not attempting to develop a search engine for global or domain-specific use. Rather our goal is to provide a portable and user-friendly framework for retrieval services by using an existing search engine. The term *system* is equivalent to service, but is used when we are emphasizing internal structure.

The data a user inputs to initiate retrieval are referred to as a *search term*. A search term may be two or more words separated by a space. The smallest unit of information to be searched is called a *record*. The scale of the service is often represented by the number of registered records. We use *document* as a synonym of record.

Despite the existence and wide adoption of internet retrieval services, such as Google, retrieval services are receiving increasing attention. Local services have significant advantages; the most promising benefit of a local full-text search system is remaining confidentiality of in-house documents and documents where intellectual property rights are in dispute. Another advantage is the ability to recognize and increase the value of the material through Web mining and access log analysis [4] [5].

Hyper Estraier, adopted in our service, is a nonresident search engine. The service invokes an executable file every time it registers a record or sends a search query. We have ensured that the service developed previously worked well. Another sophisticated search engine is Apache Lucene [6], which is written entirely in Java.

III. Related Works and Software

A number of researchers have attempted to develop digital library retrieval services for documents written using Latin characters [7] [8]. In addition, some researchers have proposed systems for other writing systems. Chen, Hsiang, Tu, and Wu [9] reported a full-text search system for Taiwanese historical documents written in Chinese. The documents were housed in the National Taiwan University Library. Alshuhri [10] attempted to index and retrieve Arabic manuscripts stored in digital libraries. Batjargal, Khaltarkhuu, Kimura, and Maeda [11] focused on traditional Mongolian scripts and provided a query method using modern Mongolian. The authors have also experimented with a database of Chinese characters in Buddhist canons [3].

There are several methods, other than search engines, to retrieve documents. We will describe these briefly and examine their disadvantages. Grep is a well-known sequential search command originally developed for Unix environments, which is now available in Windows as well. Typically, grep does not create or readily use an index. Consequently, grep retrieval time is directly proportional to the volume of content searched. In addition, the file access (open, readout, and close) time would be a consideration. For these reasons, a grep-based approach to searching vast numbers of documents would be impractical.

It is difficult to perform a multiword search on large plain text files or large Excel spreadsheets. For example, it is very hard to find documents in which two words appear in no particular order. Moreover, AND/OR/NOT searches might be ineffective on such large files. The proposed method is intended to be suitable for text search of many small files and/or a few very large files. The problem of word order can be resolved using a search engine.

Recently, with the popularization of cloud computing, document management systems are being used more frequently. Evernote, a leading service, manages user records with tags called notebooks. Evernote and other popular services can operate across multiple devices such as PCs and smartphones. However, there are some practical problems associated with these services; large volumes of data may take a long time to upload and there are terms-of-use restrictions. Another reason for researchers to hesitate to use these services is the psychological resistance to the commission of their own data.

Before search engines matured, retrieval systems were constructed using DBMS, and queries using Structured Query Language (SQL) were executed for retrieval. It is true that some DBMSs included full-text search features, but they were not suitable for searching text data written in Japanese or Chinese. In addition, these DBMS-based retrieval systems do not handle records containing multiple languages effectively. For example, the SQL Server full-text search function that runs on Windows requires a specific language to be identified to generate the index.

In general, when constructing a retrieval service based on a DBMS, it is necessary to design the database carefully in advance, considering the content and search parameters. Therefore such a retrieval service is both restricted and inconvenient. DBMS-based services are suitable for narrow data retrieval, such as dates or numbers, but not for full-text searches.

IV. The System

Before describing the main body of the system, we clarify the requirements of our portable full-text search system. The search capability requires immediate and complete retrieval without communication with other computers. Note that search noise or false positive results can be corrected by better search terms. The system must be user friendly; users can perform and control a search with their own hardware. It is desirable to require minimal steps to facilitate a search. While the system described in this paper is compatible to the previous system [1] with regard to search capability, the notable feature of the proposed system is the use of a USB flash drive with any computer to perform full-text searches.

 ISBN: 978-1-61208-280-6

It is assumed that a system user is neither familiar with nor interested in the internal structure of the system. However, he/she can perform basic operations such as Web access and installing software. With regard to ease of installation, a retrieval service using virtualization, such as VMware or VirtualBox, is not a good solution. The installation and management of a virtual OS is complicated and potentially expensive. In addition, the software might perform poorly on portable, low-performance PCs.

Even though retrieval services such as those used with iPod [12] and electronic dictionaries are attracting or extensible, we excluded these implementations because they would be ineffective for handling record registration.

A. Software Configuration

Fig. 1 shows the overview of our system. The system is stored on a USB flash drive, and includes the server programs, the full-text search executable, the administrative files for the database, the source files for system operation, and the index and document files.

The server programs used in this system are from a portable edition of XAMPP server stack. We used the Apache web server, MySQL database, and Hypertext Preprocessor (PHP) interpreter. The full-text search executable, "estcmd.exe," performs retrieval on demand. The administrative file for the database maintains the search term conversion table, the list of keywords used in notification, and the bookmarks. The system does not require user authentication.

The system source files consist of Hypertext Markup Language files and PHP files based on Smarty, a PHP template engine. System directories are arranged according to Smarty conventions, and the source files are located in respective directories. We employed Cascading Style Sheets to manipulate the appearance of the pages and JavaScript to run client-side scripts. Windows batch files were used to activate and terminate the system.

The index is read or updated by the full-text search executable during registration and retrieval. The document files are linked on the retrieval results page for easy access. The user can also open the files immediately using the directory hierarchy of the flash drive.

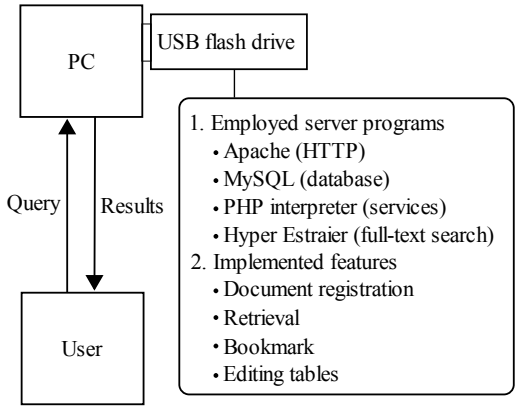


Figure 1. System overview.

B. System Setup

Fig. 2 shows the system setup and use workflows including USB flash setup.

To use the system, the USB flash drive must be prepared beforehand. To do so, we first plug the drive into a PC and install XAMPP. Then, we put Hyper Estraier executable and the system source files in place.

The database must be initialized during setup. When the batch file is run, it resolves the USB device drive letter and invokes XAMPP. We can then find the control panels and launch MySQL. We do not have to activate Apache at this time. The database is initialized by SQL setup queries.

To use the prepared drive with another PC, we stop the MySQL process, terminate XAMPP, and detach the flash drive.

The service begins by plugging the device into a PC and executing the batch file described above. Here we activate Apache and MySQL using the XAMPP control panel. When opening the Uniform Resource Locator as prescribed, including "localhost," we can see the service's front page (Fig. 3). Each page has links to retrieval ("[Search]"), document registration ("[Register]"), bookmarks catalog ("[Bookmark]"), conversion table editing ("[Term+]"), keyword editing ("[Keyword]"), and general service information ("[About]").

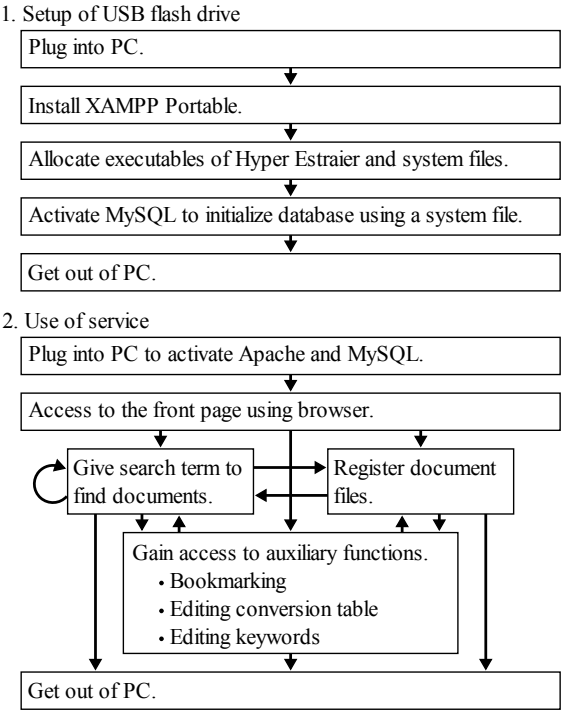


Figure 2. Setup and use workflows.

 ISBN: 978-1-61208-280-6

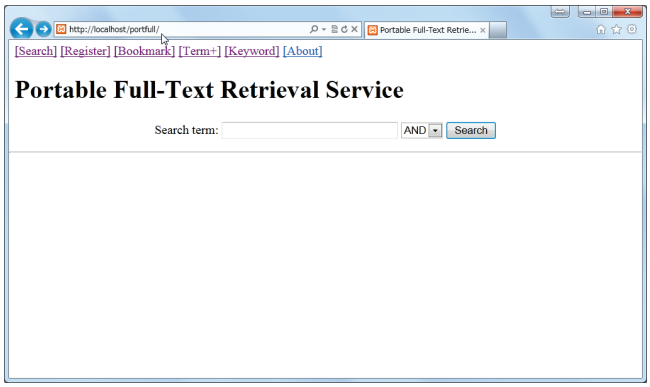


Figure 3. Front page of the service.

C. Registration of Documents

We also developed interfaces to support document registration (Fig. 4).

After opening the registration page, the user specifies the name of a folder and clicks the "Update index" button. Then, the page goes dark, and an "under construction" message appears at the center until registration finishes. At this stage, the user cannot interact with the page.

During this time, the system searches for files that should be registered by scanning the folder specified by the user. The text data are extracted from each file, and the search engine updates the index according to the data and file name.

The system accepts "txt," "doc," "docx," "pdf," "xls," and "xlsx" file extensions. There are three ways to preprocess the files for indexing. The text data of a "txt" file can be obtained through file loading. For a "doc," "docx," or "pdf" file, the batch file "estxfilt.bat" invokes "xdoc2txt.exe" to extract the content. Note that estxfilt.bat and xdoc2txt.exe are Hyper Estraier executable files. For "xls" or "xlsx" files, xdoc2txt.exe is called by the PHP interpreter to extract the data. Each row of a spreadsheet is regarded as a record to be registered. This means that an Excel file is generally split into a number of records for the indexer. This feature was derived from a request by humanities researchers who used a prototype service.

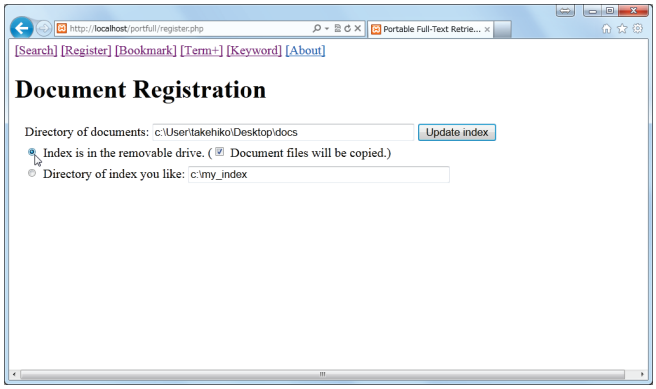


Figure 4. Front page of the service.

D. Retrieval

Fig. 5 shows the retrieval results where "1234" was given as the search term after plain text files were registered.

The screen displays the search time and the number of relevant and total records. The relevant records are shown in a table. The table also includes hyperlink to the relevant document and a snippet, i.e., content from the document around the search term. In the snippets, when the search term is highlighted, the keywords are hyperlinks. When one of these hyperlinks is clicked, a search using this keyword begins.

When many records match the search term, the table only displays the top 10 records. A "[More]" hyperlink hidden under the table is available to display the next 10 records. This visualization method avoids pagination and the page transitions. Note that search results can be reduced and improved by adding an extra word to the search term.

E. Other features

Our system supplies three reusability features, i.e., bookmarking, conversion table editing, and keyword editing.

We implemented a unique bookmark feature to allow the user to store the desired results. Note that this feature differs significantly from browser's bookmark feature; a traditional bookmark is associated with a page corresponding to a search result or a record. In contrast, the implemented bookmark module manages and displays a suite of records that is configurable by the user.

Bookmarking begins with the search result. The user checks one or more records and then clicks the bookmark button on the left of the table. The user is then presented with a new page.

The new page is composed of forms to assign bookmark names, associated notes, and links to the stored records. When the user fills in the forms and presses the registration button, the set of bookmarks is stored in the database. These bookmarks can be seen at any time. When the user selects a bookmark from the list in the top menu, the bookmark page, including the name, notes, and links to the records, is displayed.

Figure 5. Retrieval results.

Prior to describing the search term conversion table editing feature, an explanation of the search term conversion in the previous system, which was intended to support humanities researchers, is required. When a user specifies "1234" as the search term in the previous and proposed systems, the service identifies several Chinese strings, each of which means the year 1234 in Japanese. The actual Hyper Estraier query is performed by disjunctive search, where the search term consists of the original search term and derived words with the pipe symbol ("|"), i.e., the OR operator, inserted. The resulting records include at least one divisional search term.

To date, data for search term conversion have been extracted primarily from documents obtained from humanities researchers. Taking into consideration the variety of users and use cases, we must satisfy the requirement to add, modify, and remove conversion rules. To address this, we implemented an interface for editing the search term conversion table using PHP.

When the user clicks "[Term+]" to open the configuration page, he/she must first specify the candidate search term to be converted. Then he/she must click the search button. If the term has already been registered, or one or more ex-post terms exist, then the table displays all available patterns (Fig. 6), and the user can modify and/or delete these items. If the term was not registered, then the user can fill in an ex-post term and click the button for registration. Since there are currently only a few rules, the system attempts to match search term conversion to an exhaustive search of the conversion table, which is managed by the DBMS.

In a similar manner, we also provided an interface for editing keywords displayed in the search results.

V. System Evaluation

We conducted experiments to verify the usefulness of our service. Throughout experiments, we used XAMPP Portable Lite 1.8.1 (including Apache 2.4.2, MySQL 5.5.27, and PHP 5.4.7) and Hyper Estraier for Windows 1.4.10. These program files were put on a USB 3.0 flash drive.

First, we registered over 12,000 text files that were used in an evaluation of the previous system. The text files and index were stored on the PC but not on the USB flash drive.

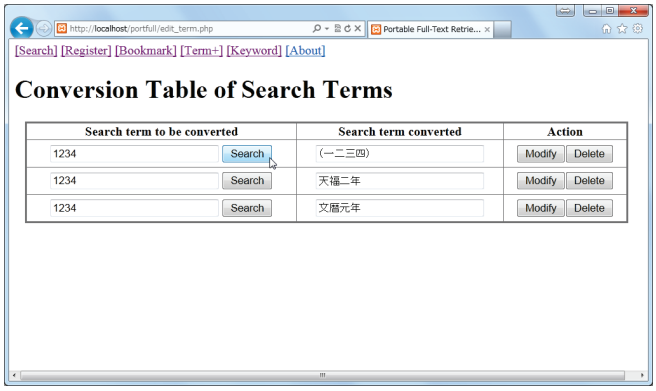


Figure 6. Editing search term conversion table.

Although it took approximately 20 min to set up, we obtained immediate search results in each case. We also ensured that when we terminated the programs and unplugged the USB flash drive, no processes or files related to the service continued to run.

We also investigated the quantitative properties of the system through index construction and search, using three PCs and two datasets. One dataset consisted of 70 PDF files (1.2 GB total), which were scanned from research papers, reports, and books written in Japanese on humanities researches. The other dataset was composed of 14 Excel files (17 MB total) with bibliographical data on ancient Japanese documents, written in Chinese and Japanese. The aggregated number of records from the Excel files exceeded 44,000. All the contents and index were put on a USB flash drive.

Table I shows indexing and search times, and provides specifications of the three PCs. Each search time in the table is the average of search times, which Hyper Estraier reported, for retrieval using predetermined three search terms. The indexing time depends on the volume of records and PC performance, although the number of records did not seem to significantly impact the performance. The type of OS and PC capability also did not have significant impact on search time.

VI. Discussion

Through the experiments, we confirmed that the system can produce a result and list of files relevant to the search term in a short time, even with a low-performance laptop PC.

The major difference between the proposed system and our previous system is the storage location of the programs, index and content. The proposed system can preserve all files on a USB flash drive. Once file registration and indexing is performed on a high-performance PC, the user can put the flash drive in a low-performance laptop PC to enjoy effective search capabilities outside of a laboratory environment. We supplied a way to use the system taking usability and security into consideration; document files and the index file can be maintained on the PC, and the user may select the file location.

TABLE I. Indexing and Search Time

Hardware	PC #1	PC #2	PC #3
OS type	Windows 8	Windows 7	Windows XP
CPU	Core i7 2.00 GHz 4 threads	Core i7 2.80 GHz 4 threads	Atom 1.60 GHz 2 threads
RAM	8 GB	8 GB	1 GB
USB Support	3.0	3.0	2.0
Dataset #1 (PDF files)			
Indexing time	650 s	704 s	1,508 s
Search time	1.20 s	1.28 s	1.07 s
Dataset #2 (Excel files)			
Indexing time	123 s	98 s	286 s
Search time	1.25 s	1.31 s	1.16 s

 ISBN: 978-1-61208-280-6

We also investigated the security and convenience of the full-text service. The operation using a client-server model was exposed to eavesdropping and access by the general public, both of which are unwanted by the content holder. Such a system would have a serious defect if it is unavailable in internet-incapable environments. In contrast, stand-alone execution has no security or operability risks; however, the computer should be protected against theft, because even though we do not need to worry about communication issues when using our system, we must consider the fact that USB flash drives are easy to lose or steal. To address this problem, security will be enhanced by encrypting the data on the flash drive and/or by employing software authentication. However, the convenience of interoperability among different computers remains a significant benefit.

Taking these properties into account, we would like to highlight the usefulness of our service. For example, when a research project task requires that text data be produced from a series of historical material, it is not always efficient to introduce and maintain a server in the laboratory or the location in which the collection is stored. Such a solution would require hardware and software costs, and must have guaranteed communication infrastructure. Since our service is free from communication requirements, the user can interact with only the USB flash drive and their own PC. Therefore it is much easier to create a text management environment that includes referencing and uploading of data.

We can register and search documents written in Japanese, Chinese, or any other language using our system because Hyper Estraier accepts UTF-8 encoded text data. For a string found in the historical material with characters unregistered with Unicode, the character code may be found in Mojikyo and described using ASCII characters in the record. In that case, such a character will not be displayed in the browser, but could be visualized by rendering software not included in our system or may visualized by the researcher.

VII. CONCLUSION

We have reported a portable full-text retrieval service that does not require communication with other computers. A simple USB flash drive can retain all the executable files, system files, document files, and the index; therefore, users can easily enjoy the service using their favorite Windows PC.

In the future, we will conduct tests to ensure that the system functions in a practical environment. In addition, an examination of the system's scalability in relation to the number of registered files or total number of characters is also required.

ACKNOWLEDGMENT

The authors thank Prof. Keigo Utsunomiya for his advice relating to humanities research and the significance of personal or small-scale document management. They would also like to thank Enago (www.enago.jp) for the English language review.

REFERENCES

[1] T. Murakawa, Y. Watagami, K. Utsunomiya, and M. Nakagawa, "Full-text retrieval system for humanities researches," Proc. Tenth Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2012), 2012, pp. 118–127.

[2] T. Tanaka, K. Fukimbara, K. Utsunomiya, H. Morikawa, and M. Nakagawa, "Database of Japanese Buddhists in the 10-13 centuries," Proc. 36th International Conference on Modelling and Simulation of Systems (MOSIS'02), 2002, pp. 265–272.

[3] T. Tanaka, Y. Nino, R. Zhang, M. Rolland, M. Nakagawa, S. Aoki, K. Utsunomiya, and T. Ochiai, "A database system of Buddhist canons," Proc. Seventh Joint Conference on Knowledge-Based Software Engineering (JCKBSE 2006), 2006, pp. 327–336.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web," Proc. 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97), 1997, pp. 558–567.

[5] Z. Su, Q. Yang, H. Zhang, X. Xu, and Y. Hu, "Correlation-based document clustering using web logs," Proc. 34th Hawaii International Conference On System Sciences (HICSS-34), 2001, pp. 3–6.

[6] H. Li, W. Li, G. Wang, and X. Peng, "Information retrieval services based on Lucene architecture," Information Computing and Applications, Communications in Computer and Information Science, 2012, pp. 638–645.

[7] M. Gonçalves, "Digital libraries," Modern information retrieval (Second edition), Pearson Education, 2011, pp. 711–735.

[8] G. Buchanan, S. J. Cunningham, A. Blandford, J. Rimmer, and C. Warwick, "Information seeking by humanities scholars," Proc. 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), LNCS 3652, 2005, pp. 218–229.

[9] S.-P. Chen, J. Hsiang, H.-C. Tu, and M. Wu, "On building a full-text digital library of historical documents," Proc. 10th International Conference on Asian Digital Libraries (ICADL 2007), LNCS 4822, 2007, pp. 49–60.

[10] S. S. Alshuhri, "Arabic manuscripts in a digital library context," Proc. 11th International Conference on Asia-Pacific Digital Libraries (ICADL 2008), LNCS 5362, 2008, pp. 387–393.

[11] B. Batjargal, G. Khaltarkhuu, F. Kimura, and A. Maeda, "Ancient-to-modern information retrieval for digital collections of traditional Mongolian script," Proc. 12th International Conference on Asia-Pacific Digital Libraries (ICADL 2010), LNCS 6102, 2010, pp. 25–28.

[12] D. Bainbridge, S. Jones, S. McIntosh, I. H. Witten, and M. Jones, "Beyond the client-server model: self-contained portable digital libraries," Proc. 11th International Conference on Asia-Pacific Digital Libraries (ICADL 2008), LNCS 5362, 2008, pp. 294–303.

 ISBN: 978-1-61208-280-6

Enabling End Users to Build Situational Collaborative Mashups at Runtime

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

***Abstract*—Web based collaboration gains importance in everyday life. However, users have to switch between dedicated collaboration tools that are not interoperable and do not satisfy the long tail of user needs. To overcome this, users would have to develop customized applications by themselves. Caused by a lack of support for users without programming experiences, existing approaches lack support for end user development (EUD) or cause high cognitive load leading to frustration. Therefore, we utilize the mashup paradigm to empower end users in configuring their collaboration environment independently. To address arbitrary collaboration demands, we enable to synchronize collaborative and non collaborative components during an application's runtime. In addition, users are able to share mashups as well as components, entirely or only in part, with an arbitrary number of collaboration partners. To further allow for individual user preferences, amongst others, we facilitate a semantic component description to synchronize different implementations of functionally similar applications.**

Keywords-Collaborative End User Development; Synchronous Groupware; Mashup Composition

I. INTRODUCTION

Indicated by an increasing number of solutions even the technological requirements of collaborative applications can be met by web browsers [1]. In this course, the number of demands for individualized collaboration tools increases, too. Especially within unstructured collaboration, characterized by its informal nature and non-hierarchical organization, not all of the users' demands can be anticipated during the development of groupware [2]. Hence, end users have to become developers themselves [3]. Within this, two basic challenges exist: Enabling users with no programming experiences to build individual collaboration tools and allowing for simultaneous development and usage.

To cope with these challenges, mashups are a promising approach, but neglect collaborative scenarios so far. Universal composition [4] combines arbitrary web resources spanning all application layers, i. e., data, logic and UI. Therefore, end users can build desired applications by combining existing components via, e. g., drag and drop. Combining these benefits and supporting end users to build solutions for synchronous collaboration independently, we propose our vision for collaborative mashup EUD in this paper. Based on an extensive literature review and experience with previously projects, we identified three major challenges for this approach, to whose solution we will contribute:

C1 To lower the training effort in new usage scenarios, users should be able to collaborate using their preferred applications. Therefore, all components, regardless of their implementation or built-in support for collaboration, have to be synchronizable and connectable. We have to develop a mechanism for unified handling of non collaborative and collaborative components, even if the latter include more sophisticated means for, e. g., awareness, which is insufficiently addressed by universal composition approaches. To this end, at least a unified component description has to be developed.

C2 In order to allow each user to use his desired device or select components according to his preferences, we will have to synchronize differently implemented components with identical functionalities. In addition, novel mechanisms for awareness between collaboration partners are required, taking into consideration the existence of different implementations and individually granted sharing levels for different application parts.

C3 To support unstructured collaboration, fine-grained sharing of mashup composition parts across all application layers is necessary. An appropriate rights management, especially regarding visualization and interaction metaphors for non programmers, is not provided sufficiently yet, and one of the major keys for success within the solution to be developed.

Addressing these challenges, we propose a mashup environment where users without programming experiences are supported to fulfil their individual collaboration needs. The approach is part of the EDYRA project, which utilizes recommendations [5] to enable mashup EUD. To extend the current solution for collaborative scenarios, we strive for a uniform handling of collaborative and non collaborative components, synchronization of differently implemented components with equal functionalities as well as fine-grained sharing of arbitrary application parts.

Consider the following scenario we refer to throughout the paper. Bob plans a sight seeing trip. To this end, he builds a mashup including a map, a hotel search and a sight advisor component. To contact Alice for some inspiration, he adds a facebook messenger component. While the latter offers built-in collaboration, the other components are not originally intended for collaborative use. However, Bob can share all of them for a synchronous usage with Alice. Because she uses

 ISBN: 978-1-61208-280-6

a smartphone, components will automatically be replaced with functionally identical pendants optimized for touch based devices. Bob restricts Alice's rights so that she can only view the selected location, but can not change it. To get additional information about the discussed sights, Alice inserts a Wikipedia component but keeps it private.

Before we present our envisioned approach in Section III, we give an overview of our foundations in Section II. Section IV briefly discusses related approaches. Finally, Section V summarizes the vision and outlines future work.

II. Conceptual Foundation

In the following, we give a brief overview of our conceptual basis, the CRUISe platform [4]. CRUISe provides universal composition through platform and technology independent combination of arbitrary web resources and services. Resources encapsulated as components are uniformly described using the Semantic Mashup Component Description Language (SMCDL) [6]. SMCDL covers the public component interface and non-functional properties, like price, author and input modalities. The interface consists of properties as well as parametrized operations and events. Optionally, ontology concepts can be annotated to clarify data semantics of parameters and properties as well as functional semantics of operations. In general, UI and service components are distinguished. Including all components, their state, event-based communication, and layout, a composition model declaratively describes mashup applications [4]. To enable context-aware selection of semantically compatible components according to, e. g., suitable target runtimes or user preferences, so called *Templates* are used [6]. To this end, templates are equally characterized by a component interface, but additionally include non-functional requirements for ranking candidates. With regard to the support for synchronous usage, CRUISe currently neither enables to synchronize components nor to individually share parts of the mashup with others.

III. Collaborative Mashup EUD

In this section, we present our environment for collaborative mashup creation by end users. The Mashup Runtime Environment (MRE) is shown in Figure 1. We take advantage of loosely coupled clients, server side components and access control using the server as proxy, through a centralized architecture. *Mashup Runtimes*, which will be reused, for instance, from CRUISe, act as execution container for mashups that context-sensitively retrieve components from the *Component Repository*. To enable execution on client and server, each environment has a Mashup Runtime. Thus, we execute private components only in clients' environments. Therefore, a collaborative mashup equals the union of all composition fragments from all clients and the server. To enable collaborative scenarios, Mashup Runtimes are extended by the following middle-ware functionalities.

Through the *Communication Manager* clients can exchange messages with the server. Each client and the server includes a *Coordination Manager*, which receives every message published at the corresponding part of the MRE first. Local messages are sent to the server, where the Coordination Manager orchestrates further processing steps. First, the *Concurrency Manager* ensures correct handling of concurrent messages by preserving their global order and drops redundant ones. Next, the *Access Manager* checks which of the MRE parts has the right to receive the current message with the help of a routing table, which includes users' sharing definitions. An *User Service* ensures trusted authentication and is tightly coupled with a *Context Service* that comprises a semantic user model. The *Awareness Manager* analyzes the current message and generates a command for all clients concerned to display, e. g., information about a new component. Therefore, the Awareness Manager is tightly couple with Access Manager, to ensure that only clients that are granted receive this information. The *Session Manager* handles necessary session data, like connected users or included components.

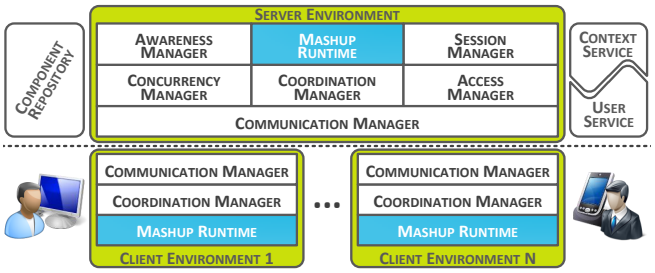


Figure 1. Architecture of the collaborative Mashup Runtime Environment

In the following, we briefly describe how we solve the identified three challenges *C1-C3* with the use of unified mashup synchronization, synchronization of heterogeneous composition fragments, and fine-grained application sharing.

A. Unified Synchronization for Mashup Components

In contrast to distinguish between collaborative and transparently adapted, non collaborative components, we strive for a hybrid approach, where users can uniformly use both kinds in one environment. Thus, regarding *C1*, users can choose components they are preferring or familiar with. Besides service or UI, components will be classified in collaborative and non collaborative. Furthermore, the former are differentiated by support for: communication, e. g. *chat*, coordination, e. g., *task list* or cooperation, e. g., *text editor*. Thereby, users can receive advice during composition extensions like, e. g., adding a communication component if only coordination and cooperation components exist. All components will be uniformly described through SMCDL.

Assisting component developers with no knowledge about collaboration support, the platform can synchronize non

 ISBN: 978-1-61208-280-6

collaborative components by using the interface definition within the SMCDL comprising properties, operations and events. Therefore, after, e. g., a marker of a map is changed, the changed property triggers an event that is captured through the corresponding client side Coordination Manager. A further client side event delegation is suppressed. Instead, the event is send within a message to the server's Coordination Manager which then, after processing the message as described above, distributes it to all Mashup Runtimes that are allowed to receive. Now, the events are propagated within the event bus of each client Mashup Runtime in parallel. Although this approach allows only for limited awareness support and basic concurrency control, its advantage is genericity: no matter which implementation the component has or which functionality it offers, it can be synchronized.

B. Synchronizing Heterogeneous Composition Fragments

To meet *C2*, first, detecting functionally similar components with different implementations has to be possible. We utilize the findings from [6]. Thereby, semantically annotated component templates are used to find components with equal interfaces. Additionally, subsumption-based matching detects differences of operations' and events' data parameters.

To provide a more expressive functionality description, we use *Capabilities*, basically consisting of an activity, e. g., *search* and an entity on which this activity is performed, e. g., *hotels*. Both are represented as ontology concepts. Based on this, we try to detect functionally equal components with different interfaces. Once two components where detected as equal, mediation techniques [6] are employed to realize synchronization. As an example, we can detect that a Google Map and a Open Street Map fulfil the functionality *location selection* and synchronize them even if one is representing current marker positions as latitude and longitude and the other as pure address string.

Enabling end users to successfully collaborate via components offering the same functionality but requiring mediation rises additional challenges concerning rights management, concurrency control and awareness. As pointed out earlier, the latter is very important for supporting end users. To offer awareness across different components, we strive for awareness ports within a component's SMCDL. These ports can be semantically annotated to express the kind of awareness information they offer, e. g., *text highlighting within a text editor*. Known semantic based mediation techniques can now be used to exchange concrete awareness information in spite of different implementations.

C. Fine-grained Release of Composition Fragments

Besides synchronizing an entire application, we enable users to share only parts of it. Facing *C3*, we facilitate to provide these parts with different restrictions to an arbitrary number of users during application runtime. To specify access privileges, so called *sharing definitions* are used characterizing three aspects: *object*, *subject*, and *permission*.

The object indicates the application part to be shared. This can potentially be any kind of composition fragment like the whole application, a couple of components, a single component, a part of a single component, or at least nothing. The subject signals the persons with whom a user wants to share an object. In principle, sharing can be divided in *private*, *shared*, or *public*. While being private only the user himself can access the object, public allows access to all individuals in the session. Shared parts are accessible for single persons or groups. How people can interact with objects can be expressed by six permissions: no right to even see the object, right to only view it, right to interact, the right to reconfigure or edit the sharing object, e. g., *add a component or change wirings between components*, and right to again create sharing definitions for others.

A collaboration session starts by initially sharing application parts with others. Within a session, every user can create an arbitrary number of sharing definitions concerning objects he has inserted or he has been granted the permission. That means, not only the session initiator can define sharing definitions. After accessing the collaborative mashup, every participant can individually extend their part of the application and individually define access rights for others. Due to different definitions affecting, e. g., the same object, conflicts can occur. Therefore, a conflict detection within the *Access Manager* checks every definition. If a conflict is detected, a visual user feedback about the conflicting definitions, the conflict reasons and proposed resolutions will be displayed.

To ensure that the proposed permission can in principle be realized with every object, components can additionally offer different *modes*. Due to the use of black box components, especially when sharing only parts of them, such modes are needed. Take a map as an example. Besides its default mode, it offers a more restrictive mode where a user only can zoom or shift the map section, but is unable to set or delete markers. Provided modes are uniformly declared in the component's semantic descriptor.

IV. Related Work

Regarding EUD of composite collaborative web applications, a wide area of previous research was reviewed.

While early groupware frameworks, like Agilo [7], try to ease groupware development through reuse of application code, due to their programmatic approach, they lack in support for users with no programming experience. To lower the development complexity, many component or widget based approaches have been presented. While Cheaib et al. [2] only focus on web services, and do not regard UI components, approaches like DyCe [8] enables to work on shared components synchronously, but offers no fine-grained right restrictions or dynamic switching between private and shared. The ROLE project [9] supports learners

 ISBN: 978-1-61208-280-6

building their own personal learning environment through combination of widgets. While users also are able to define widgets as shared or private, shared widgets are usable for every one. No adjustment during usage and no further fine-grained sharing is possible. In addition Graasp [10] distinguishes space members between owner, editor or viewer, but neither offers such fine-grained sharing definitions like we do nor a detailed concept for an end user appropriate right management. In addition, non of the solutions focus on the synchronization of heterogeneous components.

The idea to transparently synchronize encapsulated black box components by just accessing their outer interface was proposed earlier in the area of transparent adaption and collaboration transparency. Heinrich et al. [11] propose a generic approach for DOM-based rich internet applications. While they are able to synchronise arbitrary web applications, e. g., text or spread sheet editors, their solution demands identically application instances. Further, they provide no solution for end users to share only some application parts with others as well as to connect two application instances.

In the domain of mashups, only two solutions addressing synchronous collaboration were evident. Chudnovskyy et al. [12] explicitly focus on the combination of telco widgets and do not address the synchronization of arbitrary components as well as the definition of fine-grained sharings. The same holds true for Fox et al. [13], which only focus on secure collaboration and present no further concepts for sharing components through end users. Beyond this, actually no known approach for synchronously develop, share and extend mashup applications with others was proposed.

To sum up, none of the solutions provide sufficient support for the challenges identified in Section I.

V. Conclusion and Future Work

In this paper, we presented our vision for supporting end users developing individual, situational web applications for synchronous collaboration. Based on CRUISe, we address situational user needs through mashing up applications. The envisioned solution will provide three contributions. We enable to integrate and combine transparently synchronized non collaborative components as well as components explicitly supporting collaboration. In addition, to enable users with different contexts to collaborate, we synchronize differently implemented components with similar functionality using semantic component capabilities. Thirdly, we allow for fine-grained sharing of arbitrary composition fragments.

As the proposed vision is still in an early stage, further elaboration is necessary. We currently work on detailing the semantic description of collaborative components and a user study to evaluate our concepts for suitable visual interaction techniques. As a foundation we continuously expand our prototype to get user feed back.

VI. Acknowledgments

EDYRA is funded by the Free State of Saxony and the European Union within the ESF program (ESF-080951805).

References

[1] C. Gutwin, M. Lippold, and T. C. N. Graham, “Real-time groupware in the browser: Testing the performance of web-based networking,” in CSCW. ACM, 2011, pp. 167–176.

[2] N. Cheaib, S. Otmane, and M. Mallem, “Tailorable groupware design based on the 3c model,” in Int. J. Cooperative Inf. Syst., vol. 20, no. 4, 2011, pp. 405–439.

[3] T. Schümmer, “A pattern approach for end user centered groupware development,” Ph.D. dissertation, 2005.

[4] S. Pietschmann et al., “A metamodel for context-aware component-based mashup applications,” in 12th Intl. Conf. on Information Integration and Web-based Applications & Services (iiWAS). ACM, 2010, pp. 413–420.

[5] C. Radeck, A. Lorz, G. Blichmann, and K. Meißner, “Hybrid recommendation of composition knowledge for end user development of mashups,” in Proceedings of the 7th Intl. Conf. on Internet and Web Applications and Services (ICIW), 2012, pp. 30 – 33.

[6] 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, 2011, pp. 1–8.

[7] A. Guicking and T. Grasse, “A framework designed for synchronous groupware applications in heterogeneous environments,” in CRIWG, ser. Lecture Notes in Computer Science, vol. 4154. Springer, 2006, pp. 203–218.

[8] D. A. Tietze, “A framework for developing component based cooperative applications,” Ph.D. dissertation, TU Darmstadt, Sankt Augustin, 2001.

[9] S. Govaerts et al., “Towards responsive open learning environments: The role interoperability framework,” in EC-TEL, ser. Lecture Notes in Computer Science, vol. 6964. Springer, 2011, pp. 125–138.

[10] E. Bogdanov et al., “A Social Media Platform in Higher Education,” in Proceedings of the Global Engineering Education Conference (EDUCON), 2012, pp. 1–8.

[11] M. Heinrich, F. Lehmann, T. Springer, and M. Gaedke, “Exploiting single-user web applications for shared editing - a generic transformation approach,” in Proceedings of the 21st Intl. Conf. Companion on World Wide Web (WWW). ACM, 2012, pp. 1057–1066.

[12] O. Chudnovskyy et al., “End-User-Oriented Telco Mashups: The OMELETTE Approach,” in Proceedings of the 21st Intl. Conf. Companion on World Wide Web (WWW). ACM, 2012, pp. 235–238.

[13] R. Fox, J. Cooley, and M. Hauswirth, “Collaborative development of trusted mashups,” in Proceedings of the 12th Intl. Conf. on Information Integration and Web-based Applications & Services (iiWAS). ACM, 2010, pp. 255–262.

 ISBN: 978-1-61208-280-6

Issues in Conducting Expert Validation and Review and User Evaluation of the Technology Enhanced Interaction Framework and Method

Kewalin Angkananon
Electronic and Computer Science (ECS)
University of Southampton
Southampton, UK
ka3e10@ecs.soton.ac.uk

Mike Wald, Lester Gilbert
Electronic and Computer Science (ECS)
University of Southampton
Southampton, UK
mw@ecs.soton.ac.uk, lg3@ecs.soton.ac.uk

***Abstract*—A Technology Enhanced Interaction Framework has been developed to support designers and developers design and develop technology enhanced interactions for complex scenarios involving disabled people. Issues of motivation, time, and understanding when validating and evaluating the Technology Enhanced Interaction Framework were identified through a literature review and questionnaires and interviews with experts. Changes to content, system, and approach were made in order to address the identified issues. Future work will involve detailed analysis of the expert review and validation findings and the implementation of a motivating approach to user evaluation.**

Keywords - validation; expert review; user evaluation; framework; interaction

I. INTRODUCTION

This paper focuses on the issues involved with expert validation and review and user evaluation of the Technology Enhanced Interaction Framework (TEIF) and Method. The TEIF has been adapted from and extends the work of Dix [1] and Gaines [2] to support developers and designers design and develop technology enhanced interactions for complex scenarios involving disabled people. A review of interaction frameworks showed that many frameworks focus on people to people communication in the same time and at the same place but not using technology to enhance communication. Some frameworks address many interactions between humans and computers and Dix's framework for Computer Supported Cooperative Work [1] seems to address some of the possible interactions but it misses out some important interactions in the same time and at the same place situations such as people using technology to interact with real objects. In Dix's framework, the participants communicate with other participants in what is called "direct communication". Furthermore, the participants also interact with artefacts (man-made technology tools) by "controlling" or "acting". Sometimes an artefact is shared between the participants; in this case, the artefact is not only the subject of communication but can become a medium of communication, called "feedthrough". In communication about work and the artefacts of work, various means are used to refer to particular artefacts, and Dix terms this "deixis", as shown in Figure 1. However, no current framework addresses all of the interactions covered by the Technology Enhanced Interaction Framework explained in the next section. As information and communication technology has become more important in society, many researchers have been concerned with how to use technology to support communication between people and improve interactions between people, technology and objects [3] - [9]. A comprehensive review of existing frameworks [10] confirmed that there has, however, until now been no framework that has helped technology designers and developers consider all of the possible interactions that occur at the same time and in the same place. Section II briefly explains the Technology Enhanced Interaction Framework and Method. Section III describes the research methods used. Section IV presents the pilot study findings. Section V summarises conclusions and future work.

II. TECHNOLOGY ENHANCED INTERACTION FRAMEWORK AND METHOD

The Technology Enhanced Interaction Framework supports developers and designers design and develop technology enhanced interactions involving people, technology and objects and has seven main components as shown in Table I and an architecture shown in Figure 1.

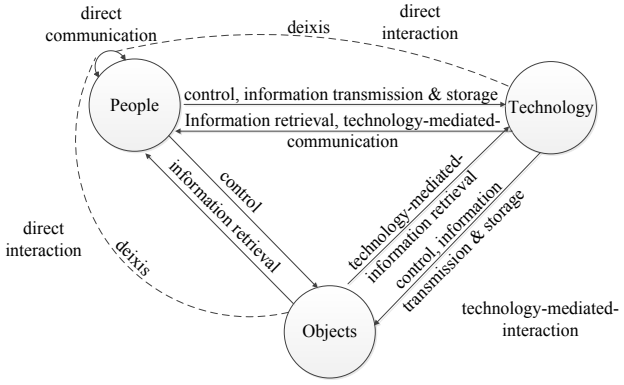


Figure 1. The Technology Enhanced Interaction Framework

The following scenario describes some problems faced by hearing impaired visitors at a museum and it is used as an example to help explain the TEIF Method by providing experts and users with requirements for an example technology solution developed using the framework. The TEIF method which has been explained in detail elsewhere [11], [12], involves 19 requirement questions based on the framework's components and a wide range of technology suggestions based on the answers to these questions.

 ISBN: 978-1-61208-280-6

TABLE I. THE TECHNOLOGY ENHANCED INTERACTION FRAMEWORK

Main Component	**Main Component of Technology Enhanced Interaction Framework**	
	Sub-component	***Example***
People	Role	A person has a role when communicating with others (e.g., presenter, audience, peer). Roles normally come in pairs such as speaker and audience (e.g., teacher and student or owner and visitor) and peer to peer (e.g., student and student or visitor and visitor).
	Ability/ Disability	People have abilities and disabilities which can affect their use of technology or understanding of language and which can lead to communication breakdown (e.g., physical, sensory, language, culture, communication, Information Technology (IT)).
Objects	Dimension	Objects have 2 dimensions (2D) or 3 dimensions (3D), and a 3D object may have a 2D representation.
	Property	Objects have colour, shape and size.
	Content	Objects have content which is human readable (text, pictures, audio, video) and machine readable (QR code, AR tag, barcode, RFID tag, NFC).
Technology	Electronic	Electronic technology has stored information, is online (e.g., internet, phone network) or offline (e.g., not connected to the internet or phone network), and is mobile (e.g., smartphone) or non-mobile (e.g., desktop computer).
	Non-electronic	Non-electronic technology is used to store information in objects (e.g., writing with a pen on paper) and is mobile (e.g., pen) or non-mobile (e.g., full-size desktop typewriter).
	User Interface	People interact with technology through its user interface (e.g., touch screen, keyboard).
	Application or Service	Electronic technology is an application (e.g., dictionary) or a service (e.g., weather forecast).
	Cost	Technology has cost (e.g., of hardware, software, maintenance).
Interactions and Communication	People-People (P-P)	People communicate verbally (speak, listen, ask, answer) and non-verbally (lip-read, smile, touch, sign, gesture, nod). When communicating, people may refer (speak or point) to particular objects or technology – this is known as deixis.
	People-Objects (P-O)	People interact with objects for two main purposes: controlling (e.g., touch, hold or move), and retrieving information (e.g., look, listen, read, in order to get information or construct personal understanding and knowledge).
	People-Technology (P-T)	People control technology (e.g., hold, move, use, type, scan, make image, press, swipe) and transmit and store information (e.g., send, save, store, search, retrieve)**.**
	People-Technology -People (P-T-P)	People use technology to transmit information to assist communication with (e.g., send sms, mms, email, chat, instant message) other people.
	People-Technology -Objects (P-T-O)	People use technology (e.g., point, move, hold, scan QR codes, scan AR tag, use camera, use compass) to transmit, store, and retrieve information (send, save, store, search, retrieve) to, in, and from objects.
Time/Place	Place	Same and different time and place yield four categories: same time (ST) and same place (SP), different time (DT) and same place (SP), different time (DT) and different place (DP), same time (ST) but different place (DP).
	Time	
Context	Location	Location affects the use of technology (e.g., indoors, outdoors). For example GPS does not work well indoors.
	Weather Condition	Weather condition may affect the use of technology (e.g., rainy, cloudy, sunny, windy, hot, cold, dry, wet). For example, the mobile phone screen doesn't work well in sunshine.
	Signal Type and Quality	Signal type can affect the quality of electronic technology (e.g., broadband, GPS, 3G, 4G).
	Background Noise	Background noise can affect the communication particularly for hearing impaired people (e.g., background music, crowded situation).
	Lighting	Light can affect the interaction (e.g., Inadequate light, too bright).
Interaction Layer	Culture	Cultural layer includes countries, traditional, language and gesture (e.g., "hello" is a normal greeting used in the culture).
	Intentionality	Intention layer involves understanding, purpose and benefit (e.g., the intent is a greeting).
	Knowledge	Knowledge layer involves facts, concepts, procedures, and principles (e.g., how to spell the word "hello").
	Action	Action layer involves actions and behaviours (e.g., pressing the correct key and not hitting neighbouring keys).
	Expression	Expression layer describes how actions are carried out (e.g., whether action is correct, accurate, prompt).
	Physical	Physical layer is the lowest layer at which people interact with the physical world (e.g., the button is depressed and so sends the electronic code for the letter to the application).

 ISBN: 978-1-61208-280-6

TABLE II. EXAMPLES OF TECHNOLOGY SUGGESTIONS

Technology suggestions	Explanation	Which requirements the technology meets															
		1a.improve communication	2a.same time/ same place	3a.presenter-audience	6b. speaker speaks Thai	7b. presenter speaks Thai	9a. hearing impaired	11a. people – people	11b. people - objects	12a.online technology	13a.mobile devices	14a.pre-prepared speech	16a. indoor	17a. noise	17e.inadequate lighting	18a. low cost solution	19a. work with smart phones	Total Score
1. Mobile web site	A Mobile Web refers to access to the world wide web, i.e. the use of browser-based Internet services, from a handheld mobile device, such as a smartphone, a feature phone or a tablet computer, connected to a mobile network or other wireless network.	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	16
2. Pre-prepared caption	Captions are text versions of the spoken word. Captions allow the content of web audio and video to be accessible to those who do not have access to audio. More information about captions see: http://webaim.org/techniques/captions/	✓	✓	✓	×	×	✓	×	✓	✓	✓	✓	✓	×	✓	✓	✓	12

Table II shows two of the technology suggestions with explanations. Although the TEIF can be applied to any disability, only one disability is mentioned in the scenario to help keep the example short and easy to understand. "Suchat Trapsin allocated some parts of his house to become the Museum of Folk Art and Shadow Puppets, in Thailand. There are exhibits of shadow puppets inside the museum, but there is no information provided in text format because Suchat normally explains the history and tradition in Thai by talking to visitors. He presents the same information in the same order every time. Chuty (who has been hearing impaired since birth) and her parents (who have some hearing loss due to their age) are local people who visit the museum. Suchat starts the talk by explaining about the exhibits. During the talk, Chuty and her parents find it very difficult to hear Suchat clearly. Chuty asks Suchat some questions about the exhibits. Suchat answers the questions, but Chuty misses some of the words. While Chuty and her parents are watching the shadow puppet show, they cannot hear the conversation clearly because of the background music, which is part of the show. It is also fairly dark which makes lip-reading very difficult for them. Suchat would like to have a technology solution that makes it easier for Chuty and her parents to understand him. There is good Wi-Fi at the museum so he would like to use Chuty's and her parents' smartphones to keep his costs low."

III. RESEARCH METHEDOLOGY

A. Pilot Study

Validation and review of the framework by experts was undertaken using an online system before engaging with the users (designers). In this study, the combination of online questionnaire on the system and interviewing were chosen. The online questionnaire gave experts time to complete the questionnaire as they could choose their preferred time and place and also could stop and return to the questionnaire whenever they wanted. Using the online questionnaire helps experts to see a prototype of the system so they can give more suggestions or comments about how to design the layout of the system. However, it might result in confusion between validating or reviewing the questionnaire and the system. Therefore, in the analysis of the results it was important to note whether the comments were about the system or the framework. For example, in the pilot test respondents gave comments about the slow response of the online system, which is not an issue about the content. The online questionnaire makes it easy to analyse the data and read the comments compared to the paper based system but doesn't help when the expert requires clarification of the questions or misunderstands some points. Therefore, the study also used interviews to discuss with the experts about any unclear information. Having constructed the questionnaire, it is important to pilot it before giving it to experts to validate and review as it is difficult even for an experienced questionnaire designer to get a questionnaire completely right the first time. Questionnaires must be piloted on a small scale sample of people characteristic of those in the survey.

To pilot the validation and review, one experienced accessibility expert and two experienced designers/ developers responded to an online questionnaire. Based on their responses changes were made to both the content and system to improve the questions, response times and layout as summarised in Table III. The pilot study participants were shown all these changes and confirmed that they were satisfied with them.

 ISBN: 978-1-61208-280-6

TABLE III. PILOT STUDY FINDINGS

Category of changes	Result of changes
Content	
Spelling and grammar mistakes	Correct and more understandable
Rewrite instructions	Clearer
Rewrite descriptions	Clearer
Add explanation of the technology suggestion tables	Help respondents understand why technologies have ticks or crosses in cells corresponding to requirements
Improve content	Make it clear and understandable without assuming knowledge
Change the image tables to html tables	Make the table accessible, now can copy the content in order to make change, can link to the websites were provided, can provide explanations in tooltip
System	
Remove the logic and always display comment box and question	System processing was slow therefore logic didn't display question before user moved on to next question and the processing icon at the top of page was out of view unless the user scrolled up
Choice, force entry to move on or just reminder	remind the respondents to provide the answer but allow blank entry

TABLE IV. THE ADVANTAGES AND DISADVANTAGES OF THREE USER EVALUATION APPROACHES

Approaches	Main Advantages	Main Disadvantages
1: Read scenario and design solution then read and understand TEIF & Self evaluate	- Less time for participants than approaches 2 and 3 - Designers may find designing more enjoyable than just reading and answering questions as in 1	- No opportunity to actually use the framework for design
2: Read scenario and design solution then read and understand TEIF then design solution again & build and get disabled person or expert understanding needs of disabled person to evaluate	- Designers may find it more enjoyable to design and develop and test and evaluate a real solution with disabled people - Developing a working technology solution and evaluating it with disabled users provides greater face validity to the evaluation	- Most time for participants as will spend much time to design and build the software
3: Read scenario A and design solution A then read & understand TEIF and suggested solution A then read scenario B and design solution B using framework and example solution design patterns (e.g., A, C, D, E) then add their solution to the patterns and Self evaluate	- Designers may find it more enjoyable and motivating and engaging than 1 or 2 by using framework with patterns to design a new solution to a new scenario. - Designers may find it more motivating than other approaches by taking part in helping their peers in designing technology and will be able to see the value of the framework for helping build a large number of patterns.	- Participants spend more time than approach 1

TABLE V. THE PROBLEMS AND POSSIBLE SOLUTIONS OF USER EVALUATION APPROACHES

Problem Type	Actual Problem	Possible Solution
Motivation	If it takes a long time to finish the task it's difficult to find the participants	- Reward (i.e., prize, put their name on published paper)
	Individual designers may get bored if just reading and answer the questions	- Get them to design because the nature of designers like designing more than reading - Inviting a group of people who have the same interest in designing and get them to interact so becomes a more interesting task - Help them to see how their work will be of value to others
Time	Individuals designing using the new framework take too much time	- Working in a team might be quicker
Understanding	Framework is difficult to understand	- Redesign the task so it helps understanding in as short a time as possible - Select participants with a good level of understanding of the task

B. Triangulation

Triangulation is a technique used to ensure the validity and credibility of the results [13] - [15] and methodological triangulation was used based on theory from existing frameworks, expert validation and review, and user evaluation. Validation is an important process particularly when an instrument is being developed to measure the construct in the context of the concepts being studied [16]. Without validation, untested data may need revision in a future study [17]. Checking reliability normally comes at the question wording and piloting stage as if an item is unreliable, then it must also lack validity [18], [19]. An expert review is a process asking the opinions, suggestions, feedback or comments from experts. For example, subject

matter experts are asked to check content of questionnaires or appropriateness of wording and terminology of items [20]. The validation of the Technology Enhanced Interaction Framework was considered by two groups: designer/developer experts and accessibility experts. The design experts focused on the main and sub-components while accessibility experts focused on checking the accessibility aspects. After the expert review and validation, user evaluation involving real users (designers) will be used to evaluate the Technology Enhanced Interaction Framework. Ryan and Deci [20] stated that there are two types of motivation: intrinsic motivation, which refers to motivation that is animated by personal enjoyment, interest, or pleasure and is usually contrasted with extrinsic motivation, which is manipulated by reinforcement contingencies. Normally, extrinsic motivations are rewards (e.g., money) for showing the desired behavior, and the threat of punishment when misbehaving. In order to engage the participants to become interested and engaged in a task which involves spending a lot of time thinking about and understanding a new idea, both intrinsic and extrinsic motivation and Interaction Design components need to be considered. An important issue that can arise when users evaluate a new idea or concept using a prototype system is that they evaluate the system rather than the idea. Using a low fidelity prototype (e.g., paper) rather than a high fidelity prototype (e.g., a functioning website) can sometimes help the user focus on the idea rather than the system. However some users may find it more difficult to evaluate the potential of an abstract concept or idea than a concrete product [21]. Possible ways in which the designers/developers might evaluate the Technology Enhanced Interaction Framework will be considered before finally deciding on the method to be used. The advantages and disadvantages of three of these possible approaches are summarized in Table IV and problems of motivation, time and understanding and their possible solutions are presented in Table V.

IV. CONCLUSION AND FUTURE WORK

Issues of motivation, time and understanding when validating and evaluating the Technology Enhanced Interaction Framework were identified through a literature review and piloting questionnaires and interviews. Changes to content, system and approach were made in order to address these issues. Future work will involve detailed analysis of expert review and validation findings and the implementation of a motivating approach to user evaluation. The updated user evaluation plans based on the analysis of the findings will also be presented at the conference.

REFERENCE

[1] A. J. Dix, "Computer supported cooperative work - a framework," Springer Verlag, pp. 23-37, 1994.

[2] B. R. Gaines, "A conceptual framework for person-computer interaction in complex systems," Systems, Man and Cybernetics, IEEE Transactions on, vol. 18, pp. 532-541, 1988.

[3] E. Berne, Games People Play – The Basic Hand Book of Transactional Analysis, New York, Ballantine Books. 1964.

[4] A. Dix, "Challenges for Cooperative Work on the Web: An Analytical Approach." Computer Supported Cooperative Work (CSCW), 6**,** 135-156, 1997.

[5] A. Dix, J. Finlay, D.G. Abowd, and R. Beale, Human-Computer Interaction, Madrid, Spain, Prentice Hall, 2004.

[6] D. Laurillard, Rethinking University Teaching: a framework for the effective use of educational technology, London, Routledge, 1993.

[7] E. Rukzio, B. Gregor, and S. Wetzstein, "The Physical Mobile Interaction Framework (PMIF)." Technical Report LMU-MI-2008-2, 2008.

[8] Y.T. Sung, K.E. Chang, H.T. Hou, and P.F. Chen, "Designing an electronic guidebook for learning engagement in a museum of history." Computers in Human Behavior, 26**,** 74-83, 2010.

[9] D. Vyas, A. Dix, and A. Nijholt, "Role of Artefacts in Mediated Communication", CHI 2008, Florence, Italy: ACM SIGCHI, 2008,

[10] K. Angkananon, M. Wald, and L. Gilbert, "Technology Enhanced Interaction Framework", 6th Annual International Conference on Computer Games, Multimedia and Allied Technology. pp 30- 35, 2013.

[11] K. Angkananon, M. Wald, and L. Gilbert, "Designing Mobile Web Solutions for Interaction Scenarios Involving Disabled People", Advances in Computer Science, 2013.

[12] K. Angkananon, M. Wald, and L. Gilbert, "Using the Technology Enhanced Interaction Framework for Interaction Scenarios involving Disabled People" 2nd International Conference on Advances in Information Technology, 2013.

[13] L. Cohen and L. Manion, Research methods in education: Routledge, 2000.

[14] H. Altrichter, A. Feldman, and P. S. Posch, Teachers investigate their work; An introduction to action research across the professions: Routledge, 2008.

[15] T. O'Donoghue and P. K. O'Donoghue, Qualitative Educational Research in Action: Doing and Reflecting: Routledge, 2003.

[16] D. F. Polit and C. T. Beck, "The content validity index: are you sure you know what's being reported? Critique and recommendations.," 2006.

[17] H. Coombes, Research Using IT. New York: Palgrave, 2001.

[18] J. Bell, Doing Your Research Project A guide for first-time researchers in education, health and social science. England: Open University Press, 2010.

[19] C. Ramirez, "Strategies for subject matter expert review in questionnaire design.," presented at the the Questionnaire Design, Evaluation, and Testing Conference, Charleston, 2002.

[20] R. M. Ryan and E. L. Deci, "Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions," Contemporary Educational Psychology, vol. 25, pp. 54-67, 2000.

[21] Y.-k. Lim, A. Pangam, S. Periyasami, and S. Aneja, "Comparative analysis of high- and low-fidelity prototypes for more valid usability evaluations of mobile devices," presented at the Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, Oslo, Norway, 2006.

 ISBN: 978-1-61208-280-6

Chord-Cube: Multiple Aspects Visualization & Navigation System for Music by Detecting Changes of Emotional Content

Tatsuki Imai
Faculty of Environment and Information Studies
Keio University
Fujisawa, Kanagawa, Japan
e-mail: t10109ti@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 presents an interactive music search-and-navigation system visualizing musical similarities based on temporal chord progression. A unique feature of this system is a 3D musical space for displaying three types of similarities in musical samples. As a fundamental feature for calculating those features, we employed chord progression in a song because chord progression is one of the most important factor in determining the overall mood of a song. For rendering the content-based relevance with a timeline structure, our system models a typical pops and rock music as a combination of the following chord progression phases: Introductive-melody, Continued-melody, and Bridge. Our 3D visualization space adopts those three chord progression as X, Y, and Z axes. Our system provides an intuitive navigation mechanism over the visualized space by putting a query song in the origin point and showing semantic distance of the inputted song and other songs. Users can utilize this 3D space to find the desired song by putting the his/her favorites song in the origin point and recognizing the semantic distance of the origin point and other songs.**

Keywords-Music; Recommendation; Visualization

I. INTRODUCTION

The change in emotion over time in a song is one of the most important factors in the selection of music to be played on modern mobile music players and smart phones. Especially, young age users will select music according to their location and mood. To support such intuitive and emotionally-based music selection, a player must provide a smart content analysis in order to extract movements of musical elements that have deep effects on human perception.

Current music database systems implemented in online music stores such as iTunes Music Store and Sony's Music Unlimited do not support such perception-oriented retrieval methods, and as users often own thousands of music in the Cloud, such situation makes users difficult to find out their desired songs intuitively even if he/she knows details of the desired music. Furthermore, owing to the temporal nature of music it is difficult to develop an effective music search environment in which users can retrieve specific music samples by using intuitive queries as searching a temporal structure requires the system to recognize the changing features of the contents in a context-dependent manner.

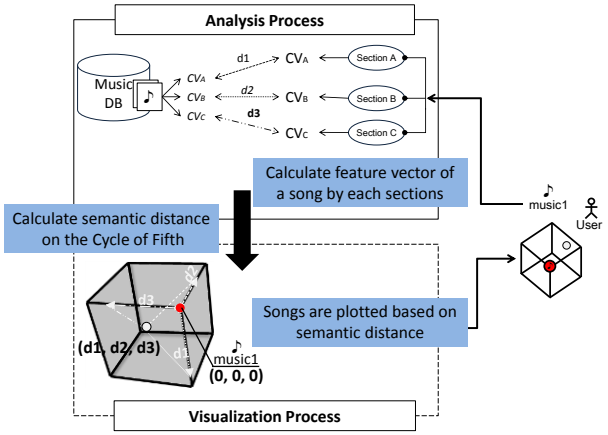


Figure 1. Conceptual diagram of Chord-Cube: Music Visualization and Navigation System within a Chord-Metric Space

Therefore, there is a need to develop a music information retrieval (MIR) method that can reflect the felt by a user as they listens to the music. Such a retrieval environment must have an interactive and navigational user interface that can visualize context-dependent relationships between songs dynamically and according to the user's viewpoint. Whereas traditional MIR systems focus on finding the most relevant song or similar songs by computing similarities or relevance according to extracted features, our system focuses on providing an integrated toolkit with which to compare song in order to create a visualization of implicit interrelationships based on emotional characteristics.

Thus, in order to detect a temporal flow of emotions instilled in a listener, we develop a stream-oriented impression analyzer of chord progression. A unique feature of our system is its "chord-vector space" in which the distance between musical chords can be calculated by analyzing the impressive behaviors of chord progression. By tracing a trajectory of chords within chord-vector space, the system can calculate represent the manner in which the music affects a listener's emotional perceptions. Our system visualizes the impressive relationship between music according to the distance calculated in this vector space.

The core concept of our visualization mechanism is a cubic metric and visual space that uses distances to represent

 ISBN: 978-1-61208-280-6

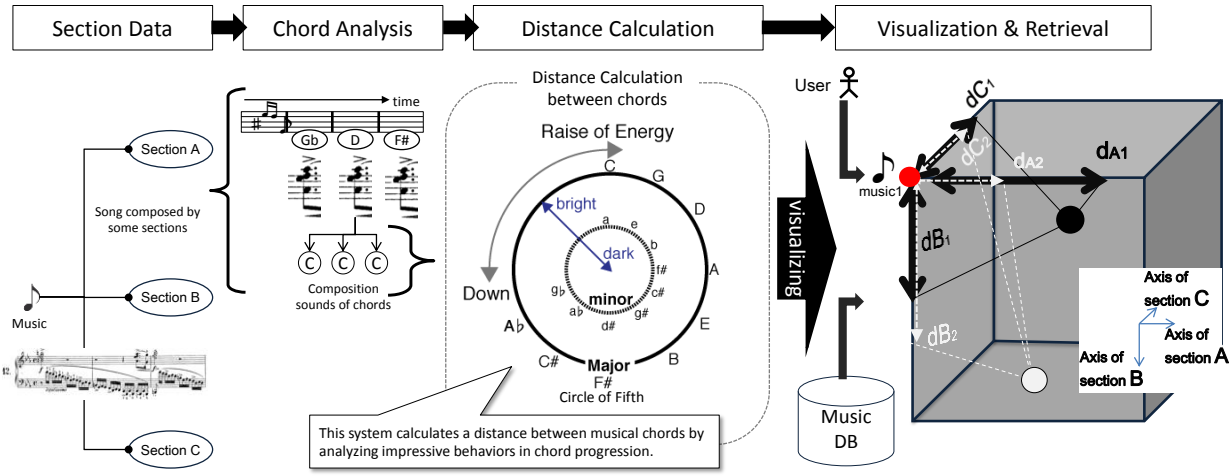


Figure 2. System Architecture of Chord-Cube

the degree of similarity between songs calculated in chord-vector-space by mapping the measured distances into a 3D graphic space for intuitive musical navigation. As shown in Figure 1, each dimension of this graphical space corresponds to the degree of similarity of chords within three respective sets of songs section types: "introductive-melody," "continued-melody," and "bridge-melody." This cube is a three-dimensional object that displays songs as points inside it. By exploiting the above chord-vector space, this system visualizes the distance between songs as a distance between points inside the cube and a vertex of cube.

Our cube accepts an initial song as an origin point in the cube. User can choose any songs as the origin point. The cube system plots other songs inside of the cube, by reflecting the distance between each song and the song at the origin point. Each axis of the cube corresponds to a musical section. For example, the x-axis corresponds to the introductive-melody section, the y-axis corresponds to the continued-melody section, and the z-axis corresponds to the bridge-melody section.

The remainder of this paper is structured as follows. Section II describes related researches of the MIR system. Section III shows an architectural overview of our system. Section IV demonstrates fundamental data structures. Section V defines core functions. Section VI shows prototype implementation of our system. Section VII performs feasibility study. Finally, Section VIII concludes this paper.

II. RELATED WORK

An MIR system utilizes many aspects of musical data; for instance, fundamental metadata such as genre and artist name, can be set as indexing keys within a conventional MIR system [1]. However, as such fundamental metadata are not sufficient to retrieve music without detailed knowledge of target data, content-based retrieval and advanced query interpretation methods are developed to find music without using fundamental metadata. In content-based music retrieval methods, a user inputs a raw music file as a query that the system analyzes and extracts several significant features from in order to identify equivalent or highly similar music samples in a database. As an example of the content-based music retrieval method, there are several input materials such as humming [2], [3], [4] and chords [5], [6] by utilizing signal frequency analysis methods [7] and power spectrum analysis methods [8]. Overall, the content-based method has advantages in terms of ease of input and the ability to generate a large amount of information reflecting musical content.

As content-based technologies are very effective in retrieving musical equivalents to inputted queries, they are widely used for copyright protection in online music sharing services. However, common users also want to be able to find new and unknown music more easily, and a method for retrieving music similar but not equal to query would be most helpful in attaining this goal. Several conventional approaches along these lines have already been made, including those by Pampalk et al. [9], who demonstrated an interface for discovering artists, Knees et al. [10], who developed a method of visually summarizing the contents of music repositories, and Stober et al. [11], who reported on an interface that can conduct music searches based on unclearly defined demands.

The most significant difference between our approach and those of the above-mentioned methods is that ours captures emotional transitions; that is, chord-vector space can capture the progression of chords as a trajectory of "how the music sounds" by representing changes of mood in music as a sequence of relevant scores corresponding to 12 types of chords. Based on this, the system can calculate the evolving distance between two chord-vectors as a continuous comparison along a timeline. Another significant innovation delivered by our method is the use of a 3D visualization space. This visualization method configures a 3D cube around an example query serving as an origin vertex point, displaying each music item according to its relevance score

 ISBN: 978-1-61208-280-6

relative to the example query. A user can operate this cube from any perspectives desired.

III. SYSTEM ARCHITECTURE

As shown in Figure 2, music navigation within the chord-cube system is achieved through integration of music content analysis and relevance visualization. As our visualization mechanism shows a dynamically measured semantic distance between music items rather than a relevance ranking, the visualized music space provides an intuitive interface for users to choose new music samples of interest.

The overall system consists of a distance calculation module and a visualization module. In order to extract chord features of a music sample, the distance calculation module inputs it as a query for analysis. The module computes distances between the chord features extracted from the query and each music item within the database based on a key technology of distance calculation that can measure the distance between two chords based on their respective temporal contexts (i.e., chord progressions). To define the relationship between chord combinations and progressions, we have developed a matrix-based data structure.

In order to make selection of desired music easy, the system displays calculated distances between samples in a 3D graphical user interface. The visualization module constructs a virtual cubic space consisting of axes corresponding to three music structures typically found in J-pop music: introductive-melody, continued-melody, and bridge. The input query is located at the origin, while target music items are located within the space according to their respective relevance scores; thus, the most relevant music item is located the closest to the origin, while irrelevant music items are scattered further away.

The system performs chord progression oriented music visualization using the following steps:

1. A user inputs a song as a criterion for finding new songs;
2. The system divides the song's chord progression into component sounds;
3. Using a method based on the cycle of fifths, the semantic distances between components are calculated and placed within a feature vector, called the chord-vector;
4. The inner products between the chord-vectors of each section are calculated to determine the similarities between each of the sections;
5. The relevance of each song is then plotted within a 3D cube in order to present an intuitive visualization of distance between the song at the vertex and the various points in the cube;
6. Further retrieval can be done by translating another song within the cube to the vertex in order to create a new relevance comparison based on the selected song as the origin.

These visualization mechanisms allow users to retrieve a desired song from an intuitive visual space based on its similarity in chord progression to the reference query song at the vertex.

TABLE I. COMPONENT SOUNDS DISTANCE MATRIX

	C	C#	D	D#	E	F	F#	G	G#	A	A#	B
C	0	0.83	0.33	0.50	0.67	0.17	1	0.17	0.67	0.50	0.33	0.83
C#	0.83	0	0.83	0.33	0.50	0.67	0.17	1	0.17	0.67	0.50	0.33
D	0.33	0.83	0	0.83	0.33	0.50	0.67	0.17	1	0.17	0.67	0.50
D#	0.50	0.33	0.83	0	0.83	0.33	0.50	0.67	0.17	1	0.17	0.67
E	0.67	0.50	0.33	0.83	0	0.83	0.33	0.50	0.67	0.17	1	0.17
F	0.17	0.67	0.50	0.33	0.83	0	0.83	0.33	0.50	0.67	0.17	1
F#	1	0.17	0.67	0.50	0.33	0.83	0	0.83	0.33	0.50	0.67	0.17
G	0.17	1	0.17	0.67	0.50	0.33	0.83	0	0.83	0.33	0.50	0.67
G#	0.67	0.17	1	0.17	0.67	0.50	0.33	0.83	0	0.83	0.33	0.50
A	0.50	0.67	0.17	1	0.17	0.67	0.50	0.33	0.83	0	0.83	0.33
A#	0.33	0.50	0.67	0.17	1	0.17	0.67	0.50	0.33	0.83	0	0.83
B	0.83	0.33	0.50	0.67	0.17	1	0.17	0.67	0.50	0.33	0.83	0

IV. DATA STRUCTURE

Our system contains four fundamental components: A) chord progression, B) component sounds distance matrix, C) Chord Vector, and D) Visualization.

A. Chord Progression

A chord progression means continuous changes of chords along a time. The system calculates the similarity of songs in terms of their respective chord progressions by using metrics in a chord-vector-space. For each song, the relevant metrics are calculated based on the semantic strengths of chords in terms of their component sounds and the occurrences of chord progression. The module divides chord progressions composed of three or more overlapping sounds into composition sounds and then calculates the correlation between occurrences of each composition sound within a song and the distances of those sounds on a cycle of fifths. The system then constructs a chord-vector space consisting of the calculated 12-dimensional values. By calculating a chord-vector based on each section of a song, comparisons between songs can be made according sectional contents. Calculating the similarity between songs is thus based on an analysis of respective chord progressions composed of three or more sounds elements in order to develop an "impression" of each song. Using a table to store the relationships between chord progression and component sounds (the chord progression component sounds table), the system is able to retrieve occurrences of various component sounds.

B. Component Sounds Distance Matrix

Component sounds distance matrix is a data matrix that stores the semantic distances between sounds on the cycle of fifths as shown in TABLE I. To calculate the similarity between songs based on their component sounds, the system uses this matrix. The values obtained by multiplying the number of occurrences of each particular sound by its respective distance represent the strength of the sounds in the song and constitute the chord vector.

C. Chord Vector

Chord-vector is based on summing the matrix consisting of the products of the semantic distance of each sound on the

 ISBN: 978-1-61208-280-6

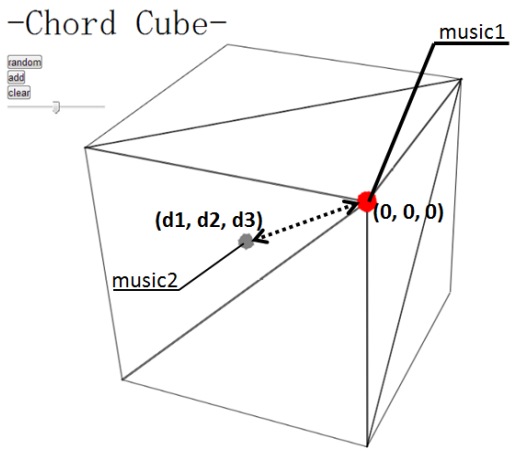


Figure 3. Over view of visualized results

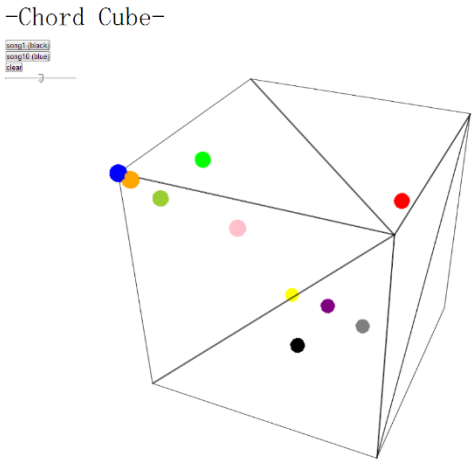


Figure 4. Implementation of the Chord-Cube. Each colored sphere represents a song.

cycle of fifths with the number of occurrences of that sound, as defined by

$$f_{CV}(d,e) \coloneqq \left(\sum_{i=1}^{12} d_{[i,1]} \cdot e_{[i]}\,,\ \sum_{i=1}^{12} d_{[i,2]} \cdot e_{[i]}\ \cdots,\ \sum_{i=1}^{12} d_{[i,12]} \cdot e_{[i]}\right) \quad (1)$$

,where *d* represents distance between the component sounds, while *e* represents number of occurrences of each component sound. The chord-vector thus generates and stores a correlation between all component sounds in each section.

D. Visualization

By using the chord-vector, the system compares a user-selected song to all songs in the music database. Defining each section of music1 (i.e., a user-imported song) as S1a, S1b, and S1c, and of music2 (another song in the database) as S2a, S2b, and S2c, the similarity calculation function distance between S1a and S2a is calculated as d1, the distance between S1b and S2b is d2, and the distance between S1c and S2c is d3. If on the 3D space consisting of the respective song section type music1 is located at the origin (0, 0, 0), then the coordinate of music2 can be represented as (d1, d2, d3); thus, the system can visualize the distances between songs as Cartesian distances in a solid body called the "chord-cube", as shown in Figure 3.

As the system is able to adopt differing user-input styles, it is able to make comparisons between songs based on varying criteria. Each song can be assigned vector values and allocated a coordinate in the cube based on its correlation to a particular criterion, creating a space that intuitively represents the semantic distance between songs in which the most relevant piece of music is located very close to the origin, while irrelevant items are more remote.

V. CORE FUNCTIONS

Chord-Vector Calculation: As mentioned in the previous section, the chord-vector matrix is derived by multiplying

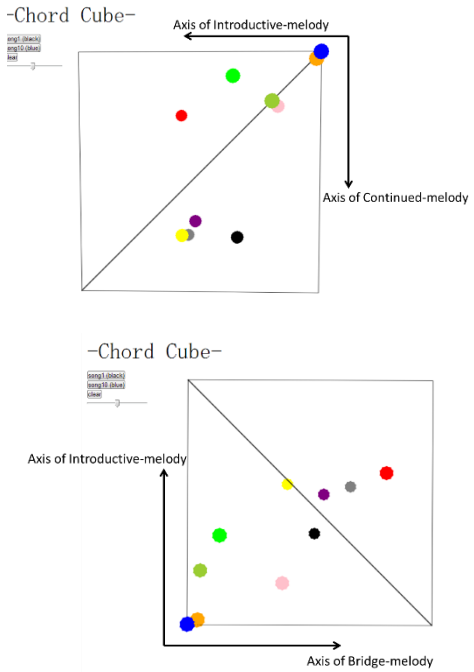


Figure 5. The system allows comparison multiple perspectives.

the component sounds distance matrix with the number of occurrences of each sound; this result consists of a 12-dimensional vector representing the strength of each sound within a section.

Chord-Vector Space: The system compares songs in terms of their representative features encoded in the 12-dimensional distance metric space ("chord-vector space") by their respective chord-vectors. Distances between sections are calculated from the inner products of vectors using

$$f_{distance}(CV_1 \cdot CV_2) \coloneqq \sum_{i=1}^{12} CV_{1[i]} \cdot CV_{2[i]} \quad (2)$$

 ISBN: 978-1-61208-280-6

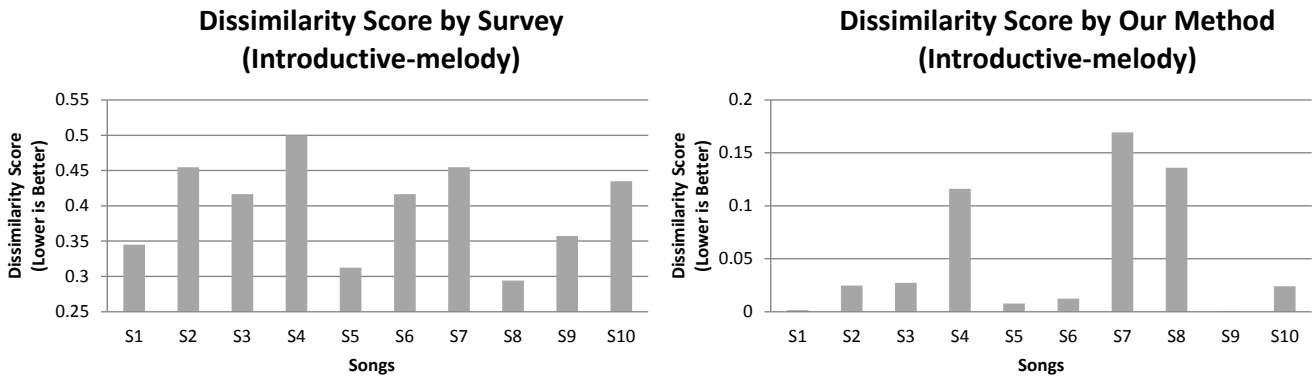


Figure 6. Results of similarity measurement for introductive-melody

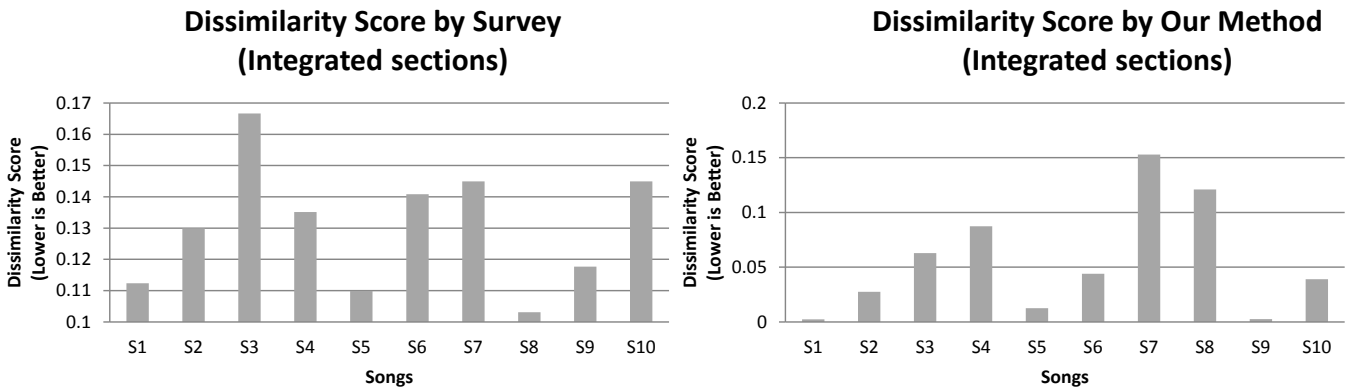


Figure 7. Results of similarity measurement for the Integrated sections

, where CV1 and CV2 are the chord-vector lengths of two different sections.

VI. System Implementation

Using Three.js Canvas and JavaScript, we implemented a prototype chord-cube system to calculate the similarity of chords by song section, as shown in Figure 4. Using JavaScript, the prototype system can represent an extended library of 3D depictions through a visualization area and an interactive user interface (UI) in which spheres and cubes can be viewed from any angle. One unique feature of this interface is that the user can see comparison outputs from any desired perspective. For example, the upper section of Figure 5 shows a perspective view from the y-axis in which the introductive-melody and continued-melody axes can be seen, representing a musical comparison between these respective sections. Similarly, the perspective in the lower figure represents a comparison between introductive-melody and bridge. As an example of the multiple perspectives viewable in the cube, the dark green and pink spheres display obvious differences between the two figures. In the upper figure, the two songs represented by these spheres have identical similarities to the blue sphere, while in lower figure they are located at differing distances. This is interpreted to mean that the two songs have are similar in terms of introductive-melody and continued-melody, and continued-melody, but difference in terms of bridge-melody.

VII. Evaluation

A. Outline of experimental studies

In this section, we evaluate the effectiveness of our system by examining its precision in providing musical analyses of input chord data. Our purpose here is to clarify the effectiveness of our method of retrieving songs by means of chord-vector space 3D visualization, and we do this by comparing the results of similarity measurements between calculated results to those of a questionnaire survey submitted to listeners who score points based on the level of similarity they feel in each section. The resulting evaluation of effectiveness comes from comparing the dissimilarities as measured by this scoring method to the distance of each song from the criterion points shown in visualization area.

For the experiment, we established one query song as a criterion and ten other songs as comparison targets. Ten listeners used a 1-to-5 scoring template to evaluate their perceptions of similarity between each comparison song and the criterion by section, and we aggregated the scoring results of each listener and converted them into reciprocal values defined as the dissimilarities by survey. We then

 ISBN: 978-1-61208-280-6

TABLE II. THE SIMILARITY RANKS OF INTRODUCTIVE-MELODY

Rank	Survey	Score	Our method	Score
1	s8	0.294118	s9	0.00028
2	s5	0.3125	s1	0.001422
3	s1	0.344828	s5	0.007712
4	s9	0.357143	s6	0.012242
5	s6	0.416667	s2	0.024543

TABLE III. THE SIMILARITY RANKS OF INTEGRATED OF SECTIONS

Rank	Survey	Score	Our Method	Score
1	s8	0.103093	s9	0.002429
2	s5	0.10989	s1	0.002381
3	s1	0.11236	s5	0.012566
4	s9	0.117647	s2	0.027525
5	s2	0.12987	s6	0.044053

used these values to calculate the distance within the chord-cube of each target song from the criterion point; this process is called collection of data dissimilarity. To evaluate the effectiveness of our method, we compared the dissimilarities by survey to the dissimilarities as calculated by the method. In experiment-1, we applied our system to measure dissimilarities of introductive-melody for each music item, while in experiment-2, we measured dissimilarities of integration of introductive-melody, continued-melody, and bridge-melody for each music item.

B. Experimental Results

Figure 6 and TABLE II shows the result of experiment-1. The left-hand side of the figure shows dissimilarity as measured by the manual survey, while the right-hand side shows dissimilarity as measured by our system. It can be seen in TABLE II that the test subjects judged songs s1, s4, s5, s8, and s9 to be highly similar to the query music, while our system retrieved songs s1, s5, s6, s9, and s10 as similar music; thus, the system correctly extracted songs s1, s5, and s9.

Figure 7 and TABLE III shows the result of experiment-2. Again, the left-hand side shows dissimilarity measured by manual survey, and the right-hand side shows dissimilarity measured by our system. By comparing Figure 6 and 7, it can be seen that the surveyed dissimilarity of song s3 drastically increases from experiment-1 to experiment-2, while our system returns identical results for all songs in both experiments. It can be concluded that our system improves its retrieval precision by integrating a differing evaluation axis into the chord-cube visualization space, and thus can effectively display multiple perspectives simultaneously.

The results for song s8, on the other hand, show that some improvements are still necessary. While the survey results judged s8 to be similar to the query music, our system judged it to be dissimilar. We believe that a perceptional gap between the theme melody and the chords progression of song s8 strongly affected the results here, as s8 has a complex chord progression but a very simple melody. However, the experimental results from the other songs closely follow the results of dissimilarity by survey, clarifying the overall effectiveness of our method for utilizing chord-metric space and 3D visualization.

VIII. CONCLUSION AND FUTURE WORKS

We proposed a music visualization and navigation system that can provide an intuitive visual retrieval method based in chord-metric space. The unique feature of this system lies in its construction of a chord-vector space to extract the transition of emotions within a song as a feature vector. In future work, we plan to improve the chord-metric space by capturing the direction of chord transitions in order to represent the change in emotional energy through the resulting motion on the cycle of fifth.

REFERENCES

[1] M. Goto and K. Hirata, "Recent studies on music information processing," Acoust. Sci. Technol., vol. 25, no. 6, pp. 419–425, November 2004.

[2] R. Type, F. Wiering, and Remco C. Veltkamp, "A Survey of Music Information Retrieval System," ISMIR 2005, pp. 153–160, 2005.

[3] A. Ghias, J. Logan, D. Chamberlin, and B.C. Smith, "Query by humming: musical information retrieval in an audio database," ACM Multimedia 95, pp. 231–236, 1995.

[4] R. B. Dannenberg, W. P. Birmingham, G. Tzanetakis, C. Meek, N.Hu, and B. Pardo, "The MUSART Testbed for Query-by-Humming Evaluation," ISMIR 2003, pp. 34–48, 2003.

[5] T. Sonoda, T. Ikenaga, K. Shimizu, and Y. Muraoka, "The Design Method of a Melody Retrieval System on Parallelized Computers," WEDELMUSIC 2002, pp. 66-73 , 2002.

[6] Heng-Tze Cheng, Yi-Husan Yang, Yu-Ching Lin, I-Bin Liao, and Homer H. Chen, "Automatic Chord Recognition For Music Classification And Retrieval," 2008 IEEE International Conference on Multimedia and Expo, pp. 1505-1508, Apil-June 2008.

[7] J. P. Bello, "Audio-based Cover Song Retrieval Using Approximate Chord Sequences: Testing Shifts, Gaps, Swaps And Beats," ISMIR 2007, pp. 239-244, 2007

[8] E. Gómez and J. Bonada, "Tonality Visualization of Polyphonic Audio," International Computer Music Conference 2005.

[9] E. Pampalk and M. Goto, "Musicrainbow: A new user interface to discover artists using audio-based similarity and web-based labeling," ISMIR 2006, pp. 367-370, 2006.

[10] P. Knees, M. Schedl, T. Pohle, and G. Widmer, "An Innovative Three-Dimensional User Interface for Exploring Music Collections Enriched with Meta-Information from the Web," 14th annual ACM international conference on Multimedia, pp. 17-24, 2006.

[11] S. Stober and A. Nürnberger, "MusicGalaxy: A Multi-focus Zoomable Interface for Multi-facet Exploration of Music Collections," CMMR 2010, pp. 273-302, June 2010

 ISBN: 978-1-61208-280-6

Compressing Large Size Files on the Web in MapReduce

Sergio De Agostino
Computer Science Department
Sapienza University
Rome, Italy
Email: deagostino@di.uniroma1.it

***Abstract*—Lempel-Ziv (LZ) techniques are the most widely used for lossless file compression. LZ compression basicly comprises two methods, called LZ1 and LZ2. The LZ1 method is the one employed by the family of Zip compressors, while the LZW compressor implements the LZ2 method, which is slightly less effective but twice faster. When the file size is large, both methods can be implemented on a distributed system guaranteeing linear speed-up, scalability and robustness. With Web computing, the MapReduce model of distributed processing is emerging as the most widely used. In this framework, we present and make a comparative analysis of different implementations of LZ compression. An alternative to standard versions of the Lempel-Ziv method is proposed as the most efficient one for large size files compression.**

Keywords-web computing; mapreduce framework; lossless compression; string factorization

I. INTRODUCTION

Lempel-Ziv (LZ) techniques are the most widely used for lossless file compression. LZ compression [1], [2], [3] is based on string factorization. Two different factorization processes exist with no memory constraints. With the first one (LZ1) [2], each factor is independent from the others since it extends by one character the longest match with a substring to its left in the input string. With the second one (LZ2) [3], each factor is instead the extension by one character of the longest match with one of the previous factors. This computational difference implies that while sliding window compression has efficient parallel algorithms [4], [5], [6], [7], LZ2 compression is hard to parallelize [8] and less effective in terms of compression. On the other hand, LZ2 is more efficient computationally than sliding window compression from a sequential point of view. This difference is maintained when the most effective bounded memory versions of Lempel-Ziv compression are considered [6], [9]. While the bounded memory version of LZ1 compression is quite straightforward, there are several heuristics for limiting the work-space of the LZ2 compression procedure in the literature. The "least recently used" strategy (LRU) is the most effective one. Hardness results inside Steve Cook's class (SC) have been proved for this approach [9], implying the likeliness of the non-inclusion of the LZ2-LRU compression method in Nick Pippenger's class (NC). Completeness results in SC have also been obtained for a relaxed version of the LRU strategy (RLRU) [9]. RLRU was shown to be as effective as LRU in [10] and, consequently, it is the most efficient one among the Lempel-Ziv techniques.

Bounding memory is very relevant with distributed processing and it is an important requirement of the MapReduce model of computation for Web computing. A formalization of this model was provided in [11], where further constraints are formulated for the number of processors, the number of iterations and the running time. However, such constraints are a necessary but not sufficient condition to guarantee a robust linear speed-up. In fact, interprocessor communication is allowed during the computational phase and experiments are needed to verify an actual speed-up. Distributed algorithms for the LZ1 and LZ2 methods approximating in practice their compression effectiveness have been realized in [6], [12], [13], where the stronger requirement of no interprocessor communication during the computational phase is satisfied. In fact, the approach to a distributed implementation in this context consists of applying the sequential procedure to blocks of data independently.

In Sections 2 and 3, we describe the Lempel-Ziv compression techniques and their bounded memory versions respectively. Section 4 sketches past work on the study of the parallel complexity of Lempel-Ziv methods leading to the idea of relaxing the least recently used strategy. In Section 5, we present the MapReduce model of computation and introduce further constraints for a robust approach to a distributed implementation of LZ compression on the Web. Section 6 makes a comparative analysis of different implementations of LZ compression in this framework and proposes an alternative to the standard versions as the most efficient one for large size files compression. Conclusions and future work are given in Section 7.

II. LEMPEL-ZIV DATA COMPRESSION

Lempel-Ziv compression is a dictionary-based technique. In fact, the factors of the string are substituted by *pointers* to copies stored in a dictionary, which are called *targets*. LZ1 (LZ2) compression is also called the sliding (dynamic) dictionary method.

Given an alphabet A and a string S in A^* the LZ1 factorization of S is $S = f_1 f_2 \cdots f_i \cdots f_k$ where f_i is the shortest substring, which does not occur previously in the prefix $f_1 f_2 \cdots f_i$ for $1 \leq i \leq k$. With such factorization, the

 ISBN: 978-1-61208-280-6

encoding of each factor leaves one character uncompressed. To avoid this, a different factorization was introduced (LZSS factorization) where f_i is the longest match with a substring occurring in the prefix $f_1 f_2 \cdots f_i$ if $f_i \neq \lambda$, otherwise f_i is the alphabet character next to $f_1 f_2 \cdots f_{i-1}$ [14]. f_i is encoded by the pointer $q_i = (d_i, \ell_i)$, where d_i is the displacement back to the copy of the factor and ℓ_i is the length of the factor (LZSS compression). If $d_i = 0$, l_i is the alphabet character. In other words the dictionary is defined by a window sliding its right end over the input string, that is, it comprises all the substrings of the prefix read so far in the computation.

The LZ2 factorization of a string S is $S = f_1 f_2 \cdots f_i \cdots f_k$ where f_i is the shortest substring, which is different from one of the previous factors. As for LZ1 the encoding of each factor leaves one character uncompressed. To avoid this a different factorization was introduced (LZW factorization) where each factor f_i is the longest match with the concatenation of a previous factor and the next character [15]. f_i is encoded by a pointer q_i to such concatenation (LZW compression). LZ2 and LZW compression can be implemented in real time by storing the dictionary with a trie data structure. Differently from LZ1 and LZSS, the dictionary is only prefix.

III. Bounded Size Dictionary Compression

The factorization processes described in the previous section are such that the number of different factors (that is, the dictionary size) grows with the string length. In practical implementations instead the dictionary size is bounded by a constant and the pointers have equal size. For LZSS (or LZ1) compression this can be simply obtained by sliding a fixed length window and by bounding the match length. Real time implementations are realized by means of hashing techniques providing a specific position in the window where a good approximation of the longest match is found on realistic data. For LZW (or LZ2) compression dictionary elements are removed by using a deletion heuristic. The deletion heuristics we describe in this section are FREEZE, RESTART, SWAP, LRU [16] and RLRU [9].

Let $d + \alpha$ be the cardinality of the fixed size dictionary where α is the cardinality of the alphabet. With the FREEZE deletion heuristic, there is a first phase of the factorization process where the dictionary is filled up and "frozen". Afterwards, the factorization continues in a "static" way using the factors of the frozen dictionary. In other words, the LZW factorization of a string S using the FREEZE deletion heuristic is $S = f_1 f_2 \cdots f_i \cdots f_k$ where f_i is the longest match with the concatenation of a previous factor f_j, with $j \leq d$, and the next character.

The shortcoming of the FREEZE heuristic is that after processing the string for a while the dictionary often becomes obsolete. A more sophisticated deletion heuristic is RESTART, which monitors the compression ratio achieved on the portion of the input string read so far and, when it starts deteriorating, restarts the factorization process. Let $f_1 f_2 \cdots f_j \cdots f_i \cdots f_k$ be such a factorization with j the highest index less than i where the restart operation happens. Then, f_j is an alphabet character and f_i is the longest match with the concatenation of a previous factor f_h, with $h \geq j$, and the next character (the restart operation removes all the elements from the dictionary but the alphabet characters). Usually, the dictionary performs well in a static way on a block long enough to learn another dictionary of the same size. This is what is done by the SWAP heuristic. When the other dictionary is filled, they swap their roles on the successive block.

The best deletion heuristic is the LRU (last recently used) strategy. The LRU deletion heuristic removes elements from the dictionay in a "continuous" way by deleting at each step of the factorization the least recently used factor, which is not a proper prefix of another one. In [9] a relaxed version (RLRU) was introduced. RLRU partitions the dictionary in p equivalence classes, so that all the elements in each class are considered to have the same "age" for the LRU strategy. RLRU turns out to be as good as LRU even when p is equal to 2 [10]. Since RLRU removes an arbitrary element from the equivalence class with the "older" elements, the two classes (when p is equal to 2) can be implemented with a couple of stacks, which makes RLRU slightly easier to implement than LRU in addition to be more space efficient. SWAP is the best heuristic among the "discrete" ones.

IV. LZ Compression on a Parallel System

LZSS (or LZ1) compression can be efficiently parallelized on a PRAM EREW [4], [5], that is, a parallel machine where processors access a shared memory without reading and writing conflicts. On the other hand, LZW (or LZ2) compression is P-complete [8] and, therefore, hard to parallelize. Decompression, instead, is parallelizable for both methods [17]. The asymmetry of the pair encoder/decoder between LZ1 and LZ2 follows from the fact that the hardness results of the LZ2/LZW encoder depend on the factorization process rather than on the coding itself.

As far as bounded size dictionary compression is concerned, the "parallel computation thesis" claims that sequential work space and parallel running time have the same order of magnitude giving theoretical underpinning to the realization of parallel algorithms for LZW compression using a deletion heuristic. However, the thesis concerns unbounded parallelism and a practical requirement for the design of a parallel algorithm is a limited number of processors. A stronger statement is that sequential logarithmic work space corresponds to parallel logarithmic running time with a polynomial number of processors. Therefore, a fixed size dictionary implies a parallel algorithm for LZW compression satisfying these constraints. Realistically, the satisfaction of these requirements is a necessary but not a sufficient

 ISBN: 978-1-61208-280-6

condition for a practical parallel algorithm since the number of processors should be linear. The SC^k-hardness and SC^k-completeness of LZ2 compression using, respectively, the LRU and RLRU deletion heuristics and a dictionary of polylogarithmic size show that it is unlikely to have a parallel complexity involving reasonable multiplicative constants [9]. In conclusion, the only practical LZW compression algorithm for a shared memory parallel system is the one using the FREEZE, RESTART or SWAP deletion heuristics. Unfortunately, the SWAP heuristic does not seem to have a parallel decoder. Since the FREEZE heuristic is not very effective in terms of compression, RESTART is a good candidate for an efficient parallel implementation of the pair encoder/decoder on a shared memory parallel system and even on a system with distributed memory. However, in the context of distributed processing of massive data with no interprocessor communication the LZW-RLRU technique turns out to be the most efficient one. We will see these arguments more in detail in the next two sections.

V. The MapReduce Model of Computation

The MapReduce programming paradigm is a sequence $P = \mu_1\rho_1 \cdots \mu_R\rho_R$ where μ_i is a mapper and ρ_i is a reducer for $1 \leq i \leq R$. First, we describe such paradigm and then discuss how to implement it on a distributed system. Distributed systems have two types of complexity, the interprocessor communication and the input-output mechanism. The input/output issue is inherent to any parallel algorithm and has standard solutions. In fact, in [11] the sequence P does not include the I/O phases and the input to μ_1 is a multiset U_0 where each element is a $(key, value)$ pair. The input to each mapper μ_i is a multiset U_{i-1} output by the reducer ρ_{i-1} for $1 < i \leq R$. Mapper μ_i is run on each pair (k, v) in U_{i-1}, mapping (k, v) to a set of new $(key, value)$ pairs. The input to reducer ρ_i is U'_i, the union of the sets output by μ_i. For each key k, ρ_i reduces the subset of pairs of U'_i with the key component equal to k to a new set of pairs with key component still equal to k. U_i is the union of these new sets.

In a distributed system implementation, a key is associated with a processor (a node in the Web). All the pairs with a given key are processed by the same node but more keys can be associated to it in order to lower the scale of the system involved. Mappers are in charge of the data distribution since they can generate new key values. On the other hand, reducers just process the data stored in the distributed memory since they output for a set of pairs with a given key another set of pairs with the same given key.

To add the I/O phases to P, we extend the sequence to $\mu_0\rho_0\mu_1\rho_1 \cdots \mu_R\rho_R\mu_{R+1}\rho_{R+1}$, where (λ, x) is the unique $(key, value)$ pair input to μ_0 with λ empty key and x input data. μ_0 distributes the data generating the multiset U_0 while ρ_0 is the identity function. Finally, μ_{R+1} maps U_R to a multiset where all the pair elements have the same key λ and ρ_{R+1} reduces such multiset to the pair (λ, y) where y is the output data.

The following complexity requirements are stated as necessary for a practical interest in [11]:

- R is polylogarithmic in the input size n;
- the number of processors (or nodes in the Web) involved is $O(n^{1-\epsilon})$ with $0 < \epsilon < 1$;
- the amount of memory available for each node is $O(n^{1-\epsilon})$;
- the running time of mappers and reducers is polynomial in n.

As mentioned in the introduction, such requirements are necessary but not sufficient to guarantee a speed-up of the computation. Obviously, the total running time of mappers and reducers cannot be higher than the sequential one and this is trivially implicit in what is stated in [11]. The non-trivial bottleneck is the communication cost of the computational phase after the distribution of the original input data among the processors and before the output of the final result. This is obviously algorithm-dependent and needs to be checked experimentally since R can be polylogarithmic in the input size. The only way to guarantee with absolute robustness a speed-up with the increasing of the number of nodes is to design distributed algorithms implementable in MapReduce with $R = 1$. Moreover, if we want the speed-up to be linear then the total running time of mappers and reducers must be $O(t(n)/n^{1-\epsilon})$ where $t(n)$ is the sequential time. These stronger requirements are satisfied by the distributed implementations of the several versions of LZ compression discussed in the next section, except for one of them, which requires $R = 2$.

VI. LZ Compression on the Web in MapReduce

We can factorize blocks of length ℓ of an input string using any of the bounded memory compression techniques with an $O(\ell)$ time, $O(n/\ell)$ processors distributed algorithm. The algorithm is suitable for a small scale system but due to its adaptiveness it works on a large scale parallel system only when the file size is large.

With the sliding window method, ℓ is equal to kw where k is a positive integer and w is the window length [6], [12], [13]. The window length is usually several thousands kilobytes. The compression tools of the Zip family, as the Unix command "gzip" for example, use a window size of at least 32K. From a practical point of view, we can apply something like the gzip procedure to a small number of input data blocks achieving a satisfying degree of compression effectiveness and obtaining the expected speed-up on a real parallel machine. Making the order of magnitude of the block length greater than the one of the window length guarantees robustness on realistic data. The window length is

 ISBN: 978-1-61208-280-6

usually several thousands kilobytes. The compression tools of the Zip family, as the Unix command "gzip" for example, use a window size of at least 32K. It follows that the block length in our parallel implementation should be about 300K and the file size should be about one third of the number of processors in megabytes.

In the MapReduce framework, we implement the distributed procedure above with a sequence $\mu_0\rho_0\mu_1\rho_1\mu_2\rho_2$ where $\mu_0\rho_0$ and $\mu_2\rho_2$ are the input and output phases, respectively. Let $X = X_1 \cdots X_m$ be the input string where X_i is a substring that has the same length $\ell \geq 300K$ for $1 \leq i \leq m$. The complexity requirements of the MapReduce model will be satisfied by the fact that ℓ is allowed to be strictly greater than 300K. The input to μ_0 is the pair $(0, X)$ mapping this element to the set S of pairs $(1, X_1) \cdots (m, X_m)$ and the reducer ρ_0 sets U_0 to S as input to μ_1. U_0 is mapped to itself by μ_1 and ρ_1 reduces (i, X_i) to (i, Y_i) where Y_i is the LZSS coding of X_i for $1 \leq i \leq m$. Finally, μ_2 maps each element (i, Y_i) of its input $U_1 = \{(1, Y_1) \cdots (m, Y_m)\}$ to $(0, Y_i)$ and ρ_2 outputs $(0, Y)$ where $Y = Y_1 \cdots Y_m$. Obviously, the stronger requirements for a linear speed-up, stated in the previous section, are satisfied by this program.

As far as LZW compression is concerned, it was originally presented with a dictionary of size 2^{12}, clearing out the dictionary as soon as it is filled up [15]. The Unix command "compress" employs a dictionary of size 2^{16} and works with the RESTART deletion heuristic. The block length needed to fill up a dictionary of this size is approximately 300K. As previously mentioned, the SWAP heuristic is the best deletion heuristic among the discrete ones. After a dictionary is filled up on a block of 300K, the SWAP heuristic shows that we can use it efficiently on a successive block of about the same dimension where a second dictionary is learned. A distributed compression algorithm employing the SWAP heuristic learns a different dictionary on every block of 300K of a partitioned string (the first block is compressed while the dictionary is learned). For the other blocks, block i is compressed statically in a second phase using the dictionary learned during the first phase on block $i - 1$. But, unfortunately, the decoder is not parallelizable since the dictionary to decompress block i is not available until the previous blocks have been decompressed. On the other hand, with RESTART we can work on a block of 600K where the second half of it is compressed statically. We wish to speed up this second phase though, since LZW compression must be kept more efficient than sliding window compression. In fact, it is well-known that sliding window compression is more effective but slower. If both methods are applied to a block of 300K and LZW has a second static phase to execute on a block of about the same length, it would no longer have the advantage of being faster. We showed how to speed up in a scalable way this second phase on a very simple tree architecture as the extended star network in [12], [18]. The idea is to factorize small sub-blocks of at least 100 bytes of the second half in parallel. This is possible with no relevant loss of compression effectiveness since the dictionary has already been learned and the factorization is static.

In the MapReduce framework, the program sequence is $\mu_0\rho_0\mu_1\rho_1\mu_2\rho_2\mu_3\rho_3$ where $\mu_0\rho_0$ and $\mu_3\rho_3$ are the input and output phases, respectively. Let $X = X_1Y_1 \cdots X_mY_m$ be the input string where X_i and Y_i are substrings having the same length $\ell \geq 300K$ for $1 \leq i \leq m$ and $Y_i = B_{i,1} \cdots B_{i,r}$ such that $B_{i,j}$ is a substring that has the same length $\ell' \geq 100$ for $1 \leq j \leq r$. The complexity requirements of the MapReduce model will be satisfied by the fact that ℓ is allowed to be strictly greater than 300K and ℓ' strictly greater than 100 bytes. Keys are pairs of positive integers. The input to μ_0 is the pair $((0, 0), X)$, which is mapped to the set S of pairs $((0, 1), X_1)$, $((1, 1), B_{1,1})$, $\cdots$, $((1, r), B_{1,r})$, $\cdots$, $((0, m), X_m))$, $((m, 1), B_{m,1})$, $\cdots$, $((m, r), B_{m,r})$ and the reducer ρ_0 sets U_0 to S as input to μ_1. U_0 is mapped to itself by μ_1. ρ_1 reduces $((0, i), X_i)$ to a set of two $(key, value)$ pairs, that is, $\{((0, i), Z_i), ((0, i), D_i)\}$, where Z_i and D_i are respectively the LZW coding of X_i and the dictionary learned during the coding process. On the other hand, $((i, j), B_{i,j})$ are reduced to themselves by ρ_1 for $1 \leq i \leq m$ and $1 \leq j \leq r$. The second iteration step $\mu_2\rho_2$ works as the identity function when applied to $((0, i), Z_i)$. μ_2 works as the identity function when applied to $((i, j), B_{i,j})$ as well. Instead, $((0, i), D_i)$ is mapped by μ_2 to $((i, j), D_i)$ for $1 \leq j \leq r$. Then, ρ_2 reduces the set $\{((i, j), B_{i,j}), ((i, j), D_i)\}$ to $((i, j), Z_{i,j})$ where $Z_{i,j})$ is the coding produced by the second phase of LZW compression with the static dictionary D_i. Finally, μ_3 maps (i, Z_i) to $((0, 0), Z_i)$ and $((i, j), Z_{i,j})$ to $((0, 0), Z_{i,j})$. Then, ρ_3 outputs $((0, 0), Z)$ where $Z = Z_1Z_{1,1} \cdots Z_{1,r} \cdots Z_mZ_{m,1} \cdots Z_{m,r}$.

The communication cost during the computational phase of the MapReduce program above is determined by μ_2. The dictionary D_i is sent from the node associated with the key $(0, i)$ to the node associated with the key (i, j) in parallel for $1 \leq i \leq m$ and $1 \leq j \leq r$. Each factor f in D_i can be represented as pc where p is the pointer to the longest proper prefix of f (an element of D_i) and c is the last character of f. Since the standard sizes for the dictionary and the alphabet are respectively 2^{16} and 256, three bytes can represent a dictionary element. Conservatively, at least ten nanoseconds are spent to send a byte between nodes. Therefore, the communication cost to send a dictionary is at least $30(2^{16})$ nanoseconds, which is about two milliseconds. Considering the fact that 300K are compressed usually in about 30 milliseconds by a Zip compressor and in about 15 milliseconds by an LZW compressor, the communication cost is acceptable.

The approach described above is not robust when the data are highly disseminated [19]. However, when compressing large size files even on a large scale system the size of the blocks distributed among the nodes is larger than 600K.

 ISBN: 978-1-61208-280-6

In order to increase robustness when the data are highly disseminated, the most appropriate approach is to apply a procedure where no static phase is involved. Therefore, new dictionary elements should be learned at every step while bounding the dictionary size by means of a deletion heuristic. It is, then, reasonable to propose LZW-RLRU as the most suitable in this context since it is the most efficient one. The relaxed version (RLRU) of the LRU heuristic is:

> RLRUp: When the dictionary is not full, label the i^{th} element added to the dictionary with the integer $\lceil i \cdot p/k \rceil$, where k is the dictionary size minus the alphabet size and $p < k$ is the number of labels. When the dictionary is full, label the $i-th$ element with p if $\lceil i \cdot p/k \rceil = \lceil (i-1)p/k \rceil$. If $\lceil i \cdot p/k \rceil > \lceil (i-1)p/k \rceil$, decrease by 1 all the labels greater or equal to 2. Then, label the $i-th$ element with p. Finally, remove one of the elements represented by a leaf with the smallest label.

In other words, RLRU works with a partition of the dictionary in p classes, sorted somehow in a fashion according to the order of insertion of the elements in the dictionary, and an arbitrary element from the oldest class with removable elements is deleted when a new element is added. Each class is implemented with a stack. Therefore, the newest element in the class of least recently used elements is removed. Observe that if RLRU worked with only one class, after the dictionary is filled up the next element added would be immediately deleted. Therefore, RLRU would work like FREEZE. But for $p = 2$, RLRU is already more sophisticated than SWAP since it removes elements in a continuous way and its compression effectiveness compares to the original LRU. Therefore, LZW-RLRU2 is the most efficient approach to compress on the Web or any other distributed system when the size of the input file is very large. In the MapReduce framework, a program sequence $\mu_0\rho_0\mu_1\rho_1\mu_2\rho_2$ implements it as the one for the LZSS compressor explained at the beginning of this section.

Decompression in MapReduce is simmetrical. To decode the compressed files on a distributed system, it is enough to use a special mark occurring in the sequence of pointers where the coding of a block ends. The input phase distributes the subsequences of pointers coding each block among the processors. If the file is encoded by an LZW compressor using a second phase with a static dictionary, a second special mark indicates for each block the end of the coding of a sub-block. The input phase distributes the coding of the first half of each block and the coding of the sub-blocks of the second half. Then, two iterations as for the compression case decompress in MapReduce. The first one decodes the first half of each block and learns the corresponding dictionary. The second sends the dictionaries to the corresponding processors for the decoding of the sub-blocks of the second half.

VII. Conclusion

We showed how to implement Lempel-Ziv data compression in the MapReduce framework for Web computing. With large size files, the robustness of the approach is preserved with scalability since no interprocessor communication is required. It follows that a linear speed-up is guaranteed during the computational phase. With arbitrary size files, scaling up the system is necessary to preserve the efficiency of LZW compression with very low communication cost if the data are not highly disseminated. The MapReduce framework allows in theory a higher degree of communication than the one employed in the procedures presented in this paper. In [11], it has been shown how the PRAM model of computation can be simulated in MapReduce under specific constraints with the theoretical framework. These constraints are satisfied by several PRAM Lempel-Ziv compression and decompression algorithms designed in the past [5], which are suitable for arbitrary size highly disseminated files. As future work, it is worth investigating experimentally if any of these algorithms can be realized with MapReduce in practice on specific files.

References

[1] A. Lempel and J. Ziv, "On the Complexity of Finite Sequences," IEEE Transactions on Information Theory, vol. 22, 1976, pp. 75-81.

[2] A. Lempel and J. Ziv, "A Universal Algorithm for Sequential Data Compression," IEEE Transactions on Information Theory, vol. 23, 1977, pp. 337-343.

[3] J. Ziv and A. Lempel, "Compression of Individual Sequences via Variable-Rate Coding," IEEE Transactions on Information Theory, vol. 24, 1978, pp. 530-536.

[4] M. Crochemore and W. Rytter, "Efficient Parallel Algorithms to Test Square-freeness and Factorize Strings," Information Processing Letters, vol. 38, 1991, pp. 57-60.

[5] S. De Agostino, "Parallelism and Dictionary-Based Data Compression," Information Sciences, vol. 135, 2001, pp. 43-56.

[6] L. Cinque, S. De Agostino and L. Lombardi, "Scalability and Communication in Parallel Low-Complexity Lossless Compression," Mathematics in Computer Science, vol. 3, 2010, pp. 391-406.

[7] S. De Agostino, "Lempel-Ziv Data Compression on Parallel and Distributed Systems," Algorithms, vol. 4, 2011, pp. 183-199.

[8] S. De Agostino, "P-complete Problems in Data Compression," Theoretical Computer Science, vol. 127, 1994, pp. 181-186.

[9] S. De Agostino and R. Silvestri, "Bounded Size Dictionary Compression: SC^k-Completeness and NC Algorithms," Information and Computation, vol. 180, 2003, pp. 101-112.

 ISBN: 978-1-61208-280-6

[10] S. De Agostino, "Bounded Size Dictionary Compression: Relaxing the LRU Deletion Heuristic," International Journal of Foundations of Computer Science, vol. 17, 2006, pp. 1273-1280.

[11] H. J. Karloff, S. Suri and S. Vassilvitskii, "A Model of Computation for MapReduce," Proc. SIAM-ACM Symposium on Discrete Algorithms (SODA 10), SIAM Press, 2010, pp. 938-948.

[12] S. De Agostino, "LZW versus Sliding Window Compression on a Distributed System: Robustness and Communication," Proc. INFOCOMP, IARIA, 2011, pp. 125-130.

[13] S. De Agostino, "Low-Complexity Lossless Compression on High Speed Networks," Proc. ICSNC, IARIA, 2012, pp. 130-135.

[14] J. A. Storer and T. G. Szimansky, "Data Compression via Textual Substitution," Journal of ACM, vol. 24, 1982, pp. 928-951.

[15] T. A. Welch, "A Technique for High-Performance Data Compression," IEEE Computer, vol. 17, 1984, pp. 8-19.

[16] J. A. Storer, Data Compression: Methods and Theory, Computer Science Press, 1988.

[17] S. De Agostino, "Almost Work-Optimal PRAM EREW Decoders of LZ-Compressed Text," Parallel Processing Letters, vol. 14, 2004, pp. 351-359.

[18] S. De Agostino, "LZW Data Compression on Large Scale and Extreme Distributed Systems," Proceedings Prague Stringology Conference, 2012, pp. 18-27.

[19] S. De Agostino, "Bounded Memory LZW Compression and Distributed Computing: A Worst Case Analysis," Proceedings Festschrift for Borivoj Melichar, 2012, pp. 1-9.

 ISBN: 978-1-61208-280-6

A Semantically Enriched E-Tendering Mechanism

Jingzhi Guo and Ka-Ieong Chan
Department of Computer and Information Science, University of Macau
Av. Padre Tomás, Pereira, S.J., Taipa, Macau
{jzguo, ma86522}@umac.mo

***Abstract*—E-tendering is widely used in corporate and government purchasing in e-business practice. Existing e-tendering platforms cannot solve semantic interoperability problems between heterogeneous e-tendering systems. This paper proposes a novel semantic-enriched e-tendering (SEET) approach, which enables heterogeneous e-tendering systems to be semantically connected and interoperable by applying collaborative conceptualization theory. Based on this approach, a SEET platform is implemented and guarantees that e-tender inviters and e-tender bidders can exchange their e-tendering documents in a semantically consistent manner.**

Keywords-electronic tenderring; tender; inviter; bid, bidder; semantic interoperability; semantic heterogeneity; semantic consistency; e-tendering document; document engineering

I. INTRODUCTION

Tendering is an important means of procurement and is intensively researched in traditional business [14]. The paper-based method of traditional tendering has been common place within the industry for a significant number of years. Electronic tendering (or e-tendering), "in its simplest form, is described as the electronic publishing, communicating, accessing, receiving and submitting of all tender related information and documentation via the internet, replacing the traditional paper-based tender processes, and achieving a more efficient and effective business process for all parties involved" [2]. It brings the convenience, saves both time and money, eliminates transcription errors, and increases speed of bid analysis to the people.

However, despite the apparent benefits of e-tendering, many companies have been slow to adopt electronic construction of tendering, and tender industry has not kept up with other industries clearly. The BCIS e-tendering survey [1] indicated only 13% participants adopt the web-based portal in e-tender while the 60% of them still use disks or other physical media to distribute tender documents. This evidenced that e-tendering is worth studying seriously.

In e-tender industry, there are many inviters and bidders that seek cooperation opportunities in various tendering platforms, where the inviters and bidders need to use tender documents to exchange tendering information. E-tender documents are often ad hoc formatted by different e-tender platforms in terms of Request for Information (RFI) [16], Request for Tender (RFT), Request for Proposal (RFP), Request for Quotation (RFQ) [10], Expression of Interest (EOI) [15], Advanced Tender Notice (ATN) [3], and Sale Tender (ST). By observation, we found that the exchanged documents are often semantically heterogeneous and highly need manual processing. Firstly, most documents used in e-tender are simply monolithic and not computer-interpretable, for example, e-mail attachment in Microsoft Word, scanned PDF or drawings in picture formats, which are only designed for human comprehension. Secondly, e-tender documents come from different platforms are not semantically consistent in concepts, because they are created in different contexts of companies and can only be locally correctly interpreted in their individual e-tender platforms. Semantic interoperability such as computer readability and computer understandability between different bidders and inviters of various e-tendering systems is key challenging issues of designing e-tendering systems. It require a feasible solution to processing inbound and outbound e-tender information.

Despite semantic interoperability problem of existing e-tendering platforms, many bidders still use heterogeneous e-tender documents to send their bids to inviters. In the situation of unavailable semantically-enriched e-tendering platforms, e-tender inviters have to accept those semantically heterogeneous bids that cannot be digitally and automatically interpreted. They have to pay high cost to manually read the bidding documents or create new forms or schema to store the received tender information that are heterogeneous, so that they can understand the received documents. These are the barriers of using existing e-tendering platforms and urgently ask for a feasible solution to the problem.

This paper aims to solve the above mentioned e-tendering problem by proposing a novel SEET approach. This approach defines a new e-tender information structure and apply collaborative concepts to construct e-tender documents, which allow e-tendering users to exchange tendering information in a semantically consistent way.

The rest of the paper is organized as follows: Section II describes the challenging research issue. Section III proposes a novel semantically-enriched e-tendering approach. Section IV discusses the related work, and finally the conclusion is made.

II. DESCRIPTION OF CHALLENGING ISSUE

Currently in industrial practice, there are at least two types of e-tender documents: one is the type of monolithic documents in plaintexts or drawings, and the other is the type of structured documents that are computer-readable but semantically heterogeneous. While believing that monolithic documents can be transformed to structured documents, a most important research issue is how an e-tendering platform can integrate heterogeneous e-tender document information

 ISBN: 978-1-61208-280-6

of different e-tendering systems in a semantically consistent manner such that semantically heterogeneous e-tender documents can be semantically interoperated.

TABLE I. BIDDER DOCUMENT XML (I) (J)

<投標書> <投標人公司> <名稱>ABC</名稱> <電話>886655</電話> </投標人公司> <投標編號>H0001</投標編號> <物品A> <數量>100</數量> <單價>8</單價> <貨幣>人民幣</貨幣> </物品A> <物品B> ……. </物品B> <交付日期>2012/05/08 </交付日期> <付款方式>現金</付款方式> </投標書>	<tenro> < p-nome> compra de matéria</p-nome> < tel> compra de matéria</tel> <não>H0001</não> <companhia>DEF</companhia> <data>2012-05-20</data> <paytype> numerário</paytype> <material A> < não >600</ não > <preço>8</preço> <Moeda>HK</Moeda> </material A> <material B> …… </material B> </tenro>

Consider an example illustrated in Table Ⅰ where Bidder A and Bidder B use e-tender XML schemas (I) and (J) to devise their e-tender documents. Respectively, an Inviter uses e-tender database schema (K), as shown in Table Ⅱ, to store received bidding information. It obvious that all of them use different languages, different terms and different schema structure. Thus, the inviter cannot understand the concepts and structures received from the bidders and it is hard to compare the bidders' e-tender documents.

TABLE II. INVITER DATABASE SCHEMA (K)

tender no	*company*	*phone*	*name*	*spec*	*unit*	*price*
H0001	*ABC*	*886655*	*glass*	….	*8*	*100*
H0002	*DEF*	….	….	….	….	
H0003	*XXX*	….	….	….	….	
H0004	…..	….	….	….	….	

This example shows that many bidders sending many e-tender documents with different schemas to the inviter, and these e-tender document schemas and used terms are unknown to the inviter. It means if bidders and inviter have no prior collaboration for schema mapping and term integration, the received bidding information cannot be interpreted and shall have the following particular problems:

- The language-different bidding information in English, Chinese and Portuguese, etc. are not computer understandable.
- The meanings of different terms, such as "名稱", "投標編號", "物品 A", "數量", "單價" and "交付日期" from Bidder A and "p-nome", "tel", "não", "companhia", "data", "paytype", "preço" from Bidder B, are not understandable by the inviter though many terms are actually equivalent to "Company", "phone" , "tenderNO", "name", "unit", " price", "Paydate" in the inviter's database.

The above problems can be summarized as a generic research problem of semantic interoperability of e-tendering systems or e-tender semantic consistency, that is, how a newly-designed e-tendering platform can semantically integrate heterogeneous e-tender information in a semantically consistent manner.

III. APPROACH TO SEMANTIC-ENRICHED E-TENDERING

Aiming at solving semantic interoperability problem of heterogeneous e-tendering systems, this paper proposes a novel SEET approach, which is a technical solution based on CONEX technology [6][7]. It provides novel methods of e-tender document representation and transfer by applying a set of semantically consistent common vocabularies of some natural languages. It is flexible for multiple bidders to work with any inviter without meaning ambiguity in interpreting heterogeneously formed e-tender documents.

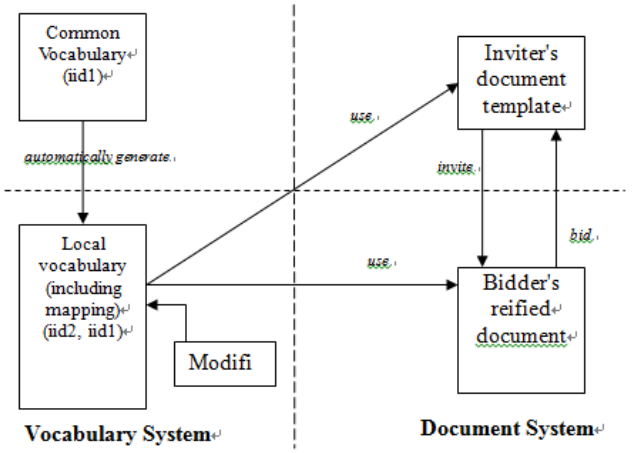


Figure 1. Seet Overview Design

Figure 1 provides an overview of SEET design. This approach divides e-tender inviters and bidders into concept designers and concept users. In vocabulary system side, designers of inviters and bidders first define their local vocabularies and then map them onto common vocabularies, which are supplied by a special vocabulary provider. In document system side, designers of inviters design their own e-tender bid templates and publish them in SEET platform. When bidders bid a tendering project, they must download corresponding e-tender bid templates and then reify them as instance bidding documents for bid submission to inviter.

SEET approach consists of the method development of SEET document representation and SEET document transfer. The former is used to create and use semantic consistent e-tender documents by applying common vocabulary and LCMap (locIid, comIid) of CONEX [6][7]. The latter is a protocol of how an e-tender document is sent, received and transformed. In the rest of this section, the methods of SEET document representation and SEET document transfer will be elaborated. Meanwhile, its theoretical foundation of semantic consistency will also be given.

 ISBN: 978-1-61208-280-6

A. SEET document representation method

SEET document is a key media to exchange e-tender information. In general, many systems are designed only for human comprehension, but lack of computer understanding. In order to achieve computer understanding, SEET develops a document representation method that separates syntax, metadata and instance data in different layers by adopting XML Product Map (XPM) [7] to develop e-tender document syntax, e-tender document templates and e-tender instance documents. It skillfully applies the hierarchical structure of XPM as document syntax and the collaborative common terms (sign of semiotics) [6] as metadata of document and individual terms for document instantiation.

XPM [7] is a sign representation language being composed of a set of terms (i.e., signs) in structure and concept (i.e., signifier and signified in semiotics [13]), which organizes the terms (or signs) in a hierarchy, such that:

$$generic\ sign = (signifier, signified) = (structure, concept) \quad (1)$$
$$atomic\ sign = (iid, term, definition, [options]) \quad (2)$$
$$complex\ sign = sign_0(sign_1, sign_2 .... sign_n) \quad (3)$$

XPM language can be used to directly describe a document or transforms a monolithic and/or non-standard document into a semantically readable and understandable document. SEET document representation by XPM can enable e-tender bid documents to be computer-readable and computer-understandable. Particularly, an XPM SEET document is represented in three separate layers: structure layer, pattern (template) layer and instance layer. By this representation separation, each layer can be arbitrarily designed and work independently.

Structure layer. It is a document structure (i.e., syntax) layer, which establishes a syntax foundation of designing and creating personalized templates for each e-tender inviter. The syntax of any personalized e-tender document template is as the same syntax of any complex sign of Formula (2), in which each atomic sign of Formula (1) is a tuple such that atomic sign = (tid, term, [interpretation], [context]), where tid is a structural sign identifier, term is a word or phrase, [interpretation] is a definition or annotation of a term, and [context] is a term reference to specify a context where a word or phrase is defined.

Pattern layer. It is a document template layer, which is based on the syntax layer and use the collaborative concepts (i.e., defined terms as atomic signs) of CONEX vocabulary [6][7] to develop personalized templates as abstract complex signs (Formula 3) according to the required e-tender information. E-tender inviters of e-tendering systems are responsible for defining different patterns (i.e., document templates) used in e-tendering systems.

Instance layer. It is a document reification layer, which instantiates document templates to reified documents as concrete/particular complex signs (Formula 3). During document instantiation, users (mostly e-tender bidders) fill in e-tender templates with local terms that are mapped onto common terms of CONEX vocabularies [6][7]. Since local terms are different in natural languages, local reified e-tender documents are semantically interoperable via common terms between different natural languages and dialects, thus maintaining semantic consistency between heterogeneous systems.

With the above-mentioned three layers, an e-tender document can be displayed in two modes: computer display mode and human display mode. The former is a mode of computer-understandable for computer to use, utilizing internal identifiers (IID) of atomic signs (Formula 2). It mainly handles computer processes when e-tender documents are transformed and stored. The latter is a mode of visual and human-readable. It displays e-tender documents in human-readable and human-understandable natural languages, utilizing terms of atomic signs (Formula 2). Table III is an example of a simplified XPM file in computer display mode and Table IV is an example of the simplified XPM file in human display mode.

TABLE III. COMPUTER DISPLAY MODE (A SIMPLIFIED XPM FILE)

```
<xpm:sign xpm:tid="21:15" xpm:refs="20017">
    <xpm:sign xpm:tid="22:21" xpm:refs="20045">
        <xpm:sign xpm:tid="23:22" xpm:refs="20020"/>
        <xpm:sign xpm:tid="24:22" xpm:refs="20021 20019" />
  </xpm:sign>
<xpm:sign xpm:tid="39:35" xpm:refs="20023 20006"/>
```

TABLE IV. HUMAN DISPLAY MODE (A SIMPLIFIED XPM FILE)

```
<xpm:sign xpm:term="Requirement">
<xpm:sign xpm:term="Object_A">
            <xpm:sign xpm:term="Quantity"/>
            <xpm:sign xpm:term="UnitPrice" />
  </xpm:sign>
<xpm:sign xpm:term="Payment Date"/>
```

B. SEET document transfer method

SEET document transfer method describes how an e-tender document is transferred from one local e-tendering system to another local e-tendering system, yet maintaining semantic consistency in the meaning of e-tender documents.

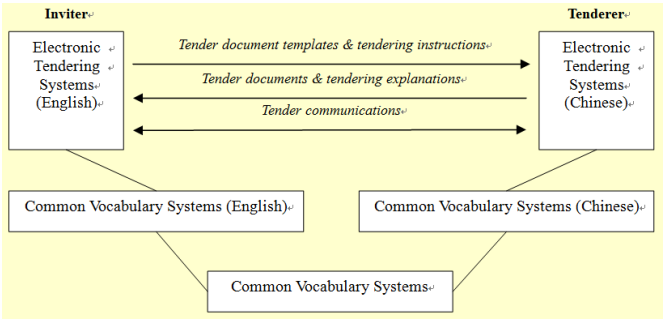


Figure 2. Seet Document Transfer Method

Figure 2 illustrates the e-tender document transfer protocol, which guarantees that the exchanged e-tender documents are semantically consistent without meaning ambiguity. It shows the exchange process of e-tender document templates and reified documents between an inviter and possibly many bidders in different natural languages.

 ISBN: 978-1-61208-280-6

SEET document transfer method provides an automatic transparent mechanism. It maintains semantic consistency of document information through mapping local terms of e-tendering documents onto terms of common vocabularies with same internal identifiers (IID) [6]. During document sending in document exchange, all terms of a local e-tender document in local natural language will first be transformed to local term IDs (locIids) based on local vocabulary, and then transformed to common term IIDs (comIids) of common vocabulary based on local-to-common term mapping datasets of LCMAP(locId, comId) at sending company. During document receiving in document exchange, all terms of a common e-tender document will first be transformed from comIids to locIids based on LCMAP (locId, comId) at receiving company, and then transformed to local terms in local natural language for human interpretation based on local vocabularies.

For example, given two vocabularies and a mapping LV_1 (L_IID: 10017→name), CV_1 (C_IID: 20012→name), and $LCMAP_1$ (10017, 20012), and two vocabularies and a mapping LV_2 (L_IID: 10011→名), CV_2 (C_IID: 20012→名稱), and $LCMAP_2$ (10011, 20012), then based on a document transfer process of $LV_1 \leftrightarrow LCMAP_1 \leftrightarrow LCMAP_2 \leftrightarrow LV_2$, we have (10017, name) ↔ (10017, 20012) ↔ (10011, 20012) ↔ (10011,名) such that "name" in e-tendering document 1 of company 1 is semantically consistent with "名" in e-tendering document of company 2 without meaning ambiguity in semantic understanding.

C. Theoretical foundation of SEET approach

SEET approach is built on a proven theoretical foundation of semantic consistency model of collaborative conceptualization theory [6], that is, structure mapability, concept equivalence and context commonality.

First, common terms of common vocabularies in different natural languages are collaboratively designed, where all common terms in different natural languages share a same common identifier (comId). This guarantees that any common vocabulary is identical in meaning interpretation and can be unambiguously used by local vocabulary designers in meaning when making local-to-common term mapping. It further guarantees semantic consistency between local terms and common terms, that it, concept equivalence.

Second, semantic consistency is not only maintained in single local-to-common term mapping but also in e-tendering document templates. This owes to the generic XPM language that is used to design e-tendering document templates such that all e-tendering document templates are designed by e-tendering inviters. E-tendering bidders only apply the e-tendering document templates to reify the templates to produce instance e-tendering documents. This guarantees that all e-tendering document processing structures are the same, that is, structure mapability onto a same hierarchy of sign IDs for executables to have same execution effects in computing.

Third, collaborative common term design, local-to-common term mapping and common e-tendering document templates provide a common context for all natural language-different e-tendering systems. Such common context guarantees that all meaning interpretations on all local receiving e-tendering documents by the receivers are exactly the same as the meaning interpretations on all local sending e-tendering documents by the senders.

Following semantic consistency model of CONEX, any local e-tendering document of one e-tendering system can be semantically represented and transferred to another e-tendering system for local interpretation without the loss of the original meaning.

IV. SEET COLLABORATION PLATFORM

Based on the SEET approach designed above, this section implements the SEET platform illustrated in Figure 3, which consists of a layer of user web-interface and a layer of concept collaboration.

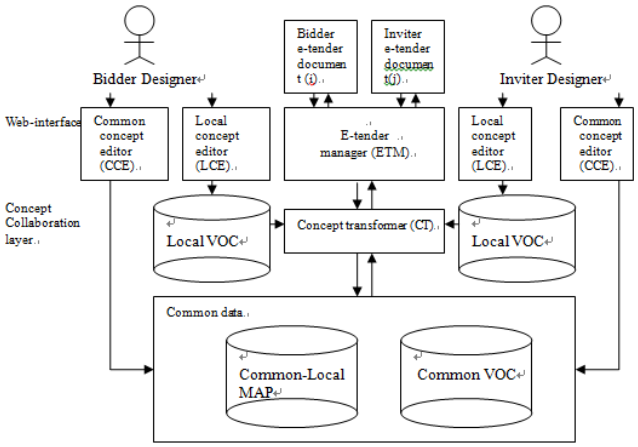


Figure 3. SEET Collaboration platform

A. Layer of User Web-Interface

The layer of user web-interface consists of concept editor and e-tender manager.

Concept Editor. It consists of a common concept editor (CCE) and a local concept editor (LCE). The former is engaged in designing common terms of the corresponding common vocabulary (common VOC). The latter is responsible for designing local terms for local vocabulary (local VOC) and builds local-to-common maps (Common-Local MAP).

E-tender Manager (ETM). It consists of document manager and tender project manager. The former controls the document representation and e-tender process. The latter manages tender project for users to be easy to read e-tendering document and observe e-tender project status.

B. Layer of Concept Collaboration

The layer of concept collaboration consists of many local vocabularies, many common vocabularies, local-to-common mapping repository and concept transformer.

Local Vocabulary (local VOC or LV). It is designed for a company to record its local term information. It is composed of local concepts written in XPM and structured in a triple, such that:

 ISBN: 978-1-61208-280-6

$$LV_{XPM} = (locId, AN, FC) \quad (4)$$

where "locId" is a unique identifier of a local concept. "AN" is the definition of the locId-ed concept. "FC" is a formal concept in the form of a word or a phrase that is defined by AN and is readable and understandable by human.

Common Vocabulary (common VOC or CV): It is designed for a natural language to record its common term information universal to all natural languages. It is formatted in XPM and structured in a triple, such that:

$$CV_{XPM} = (comId, AN, FC)^{LANG} \quad (5)$$

where "comId" is a unique internal identifier of a common concept. "LANG" indicates that AN and FC are written in natural language of LANG.

Common-Local Map repository (Common-Local MAP or CLM): It stores local-to-common mapping results that is created by local users on local concept editor. It is formatted in XPM and structured in a couple, such that:

$$CLM_{XPM} = (LocId, ComId) \quad (6)$$

where LocId is semantically equivalent to ComId.

Concept transformer (CT): It implements the SEET document transfer method that transforms one local e-tender document of one e-tendering system to another local e-tender document of another e-tendering system without losing semantic consistency.

The implementation of SEET platform assures that any local e-tender bids can be submitted to a corresponding e-tender inviter for semantically consistent interpretation without semantic ambiguity.

V. Features of SEET Platform

SEET platform presents some important features comparing with traditional paper-based tendering systems and existing e-tender websites. These features are information integration, flexibility, availability, semantic unambiguity, compatibility, extensibility and maintainability, which are shown in Table V.

TABLE V. Comparison of Paper Base, E-tender box and SEET

Features	Paper base	e-tender Box [4]	SEET
Main Technology	Paper base	Doc, xml, pdf	XPM
Information Integration	none	part	whole
Flexibility	Low	Medium	High
Cost	Low	High	Medium
Semantic unambiguity	Low	Medium	High
Compatibility	Low	Medium	High
Expansibility	Low	Medium	High
Maintainability	Low	Medium	High

SEET platform can provide these features because SEET approach is designed on a generic method of document structuring and term collaborative editing. Firstly, SEET focus on the capability of maintaining semantic consistency above e-tendering platform. It bases on CONEX technology to effectively solve two specific problems. (a) Natural language problem: many bidders and many inviters are located in different countries and regions and adopt different natural languages for bidding and invitation in e-tenders. (b) Inconsistent interpretation problem: inviters and bidders interpret incoming tendering documents differently from the interpretation of the document originators.

Second, SEET platform emphasizes on semantic unambiguity and flexibility features for both bidders and inviters such that they share same meaning for a same yet differently represented document. For a concept-same e-tendering document, it can be feasibly represented as heterogeneous documents in both forms and terms. Architecturally speaking in design, it clearly separates document syntax from document templates and separates document templates from document instance. By this separation, e-tendering documents are easy to be maintained, extended, and be compatible with existing language-different documents.

Third, SEET platform pays high attention on contexts. It represents e-tendering documents in two modes of computer display and human display. This allows contextual interpretation of e-tendering documents in heterogeneous contexts or system environments and increases compatibility of e-tendering documents in different companies.

VI. Related Work

A. Electronic Tendering

Tendering has been gradually transformed from paper-based tendering into electronic tendering. Companies of many countries have already made many efforts of building a lot of e-tendering platforms on Web (e.g., tenders.nsw.gov.au, unitender.com, and electronictender.com). However, existing e-tender platforms are far more than perfect. Most e-tendering systems combines drawings and PDF files with computer-readable documents. Some provide online tendering information to bidders together with terms and conditions. For example, e-Tender Box (ETB) [4] provides online services for suppliers to view tender notices, standard terms and conditions for tenders arranged by Government Logistics Department (GLD) as well as contract award notices of the bureau and departments of the Hong Kong Special Administrative Region Government.

Research on e-tendering focuses on interoperability among heterogeneous distributed e-tendering systems. Interoperability issue can be divided into sub-issues of mapping different e-tender database schemas, intelligent integration of e-tendering information, and electronic document exchange among different parties [9]. In another classification, e-tender interoperability problem can be discussed by the aspects of technical interoperability (refers to the ability to connect system by defining standard protocols and data formats), semantic interoperability (refers to the exchange of information in an understandable way within and across organizational borders), and organizational interoperability (refers to enabling processes to cooperate) [11]. Amongst these aspects, semantic interoperability is most challenging.

 ISBN: 978-1-61208-280-6

B. Semantic Interoperability

Semantic interoperability between distributed heterogeneous systems is one of the most important research problems in research areas of database, semantic web and electronic commerce. Earlier researches focus on the work in heterogeneous database systems (Kashyap and Sheth [8]), workflows and service oriented architectures (Nagarajan et al [12]). They explore approaches to heterogeneous database schemas integration and techniques of XML schema mediation (Gomadam et al [5]). Nevertheless, since information sources for mediation are mostly created and run in different contexts, predefined mediation rules cannot cover all heterogeneous source information, particularly the information created after the making of mediation rules. This leads to semantic conflicts between context-different distributed systems.

Collaborative concept exchange (CONEX) [6][7] is an approach to collaborative conceptualization of heterogeneous concepts of disparate information sources for consistent semantic interoperation. It introduces collaborative editing of heterogeneous concepts of different systems as the core methodology of resolving semantic conflicts. There are advantages of applying CONEX in designing e-tendering platform. It can inherit the key feature of semantic consistency between documents from CONEX approach.

VII. CONCLUSION

This paper has proposed a novel SEET approach to solving a challenging issue of semantic interoperability for e-tendering document exchange between semantically heterogeneous e-tendering systems, which consist of e-tender inviters and e-tender bidders who are located in different countries and regions and write e-tendering documents in different natural languages. SEET approach enables semantic interoperability by integrating heterogeneous e-tendering systems of e-tender inviters and e-tender bidders through novel SEET methods of e-tendering document representation and transfer based on the theory of collaborative conceptualization [6]. E-tendering document representation method has applied XML Product Map (XPM) to develop e-tender document universal syntax and utilize collaboratively designed common terms of CONEX [7] to build semantically consistent mapping between common terms and local terms and to construct e-tender document templates. E-tendering document transfer method has built a document transfer protocol that guarantees that any exchanged document is semantically consistent between document sender and document receiver. Based on SEET approach, the implemented SEET platform has several good features, such as semantic unambiguity, flexibility and compatibility, comparing with paper-based tendering systems and existing Web-based e-tendering platforms.

Contributions made in this paper are: (1) applied collaborative conceptualization theory to design semantically consistent e-tendering systems, (2) proposed a SEET design approach, (3) implemented a SEET platform of e-tendering systems.

SEET approach in this paper is still evolving. In future, more stringent work will be provided in the aspects of business model development of e-tendering, technical implementation of various patterns of e-tendering process, and evaluation matrix development on various e-tendering routine processes.

ACKNOWLEDGMENT

The research work reported in this paper has been partially supported by University of Macau Research Committee research grant number MYRG069(Y2-L2)-FST12-GJZ.

REFERENCES

[1] BCIS, "2009 e-Tendering Survey Report," RICS, 2009.

[2] S. Christensen and W. Duncan, "Maintaining the integraity of electronic tendering - reflections on the capacity of the Australian legal framework to meet this challenge," eLaw Journal, Murdoch University Electronic Journal of Law, 13(2) , 2006, pp. 8-36.

[3] Deirdre, "MIA Consulting Services. Using Advanced Tender Notices," MIA Consulting Services. 20/10/2011. http://www.miaconsultingservices.com.au/using-advanced-tender-notices.

[4] E-tender box, "Government Logistics Department (GLD)," Hong Kong. http://www.gldpcms.gov.hk. Access on 3rd Jun 2012.

[5] K. Gomadam, A. Ranabahu, L. Ramaswamy, A. Sheth and K. Verma, "Mediatability: Estimating the degree of human involvement in xml schema mediation," in: Proc. ICSC, 2008, pp. 394–401.

[6] J. Guo, "Collaborative Conceptualization: Towards a Conceptual Foundation of Interoperable Electronic Product Catalogue System Design," Enterprise Information Systems 3(1), 2009, pp.423-438.

[7] J. Guo, "Collaborative Concept Exchange," VDM Publishing, Germany, 2008.

[8] V. Kashyap and A. Sheth, "Semantic and schematic similarities between database objects: A context-based approach," VLDB J., 5(4), 1996, pp. 276–304.

[9] A. Kayed and R. Colomb, Infrastructure for Electronic Tendering Interoperability, 1999, doi:10.1.1.22.4237.

[10] S. Mhay and C. Coburn, "Request for...Procurement Processes (RFT RFQ RFP RFI)," The Negotiation Experts. http://www.negotiations.com/articles/procurement-terms/.

[11] A. Mondorf and M. Wimmer, "Interoperability in e-Tendering: The Case of the Virtual Company Dossier," In: Proc. ICEGOV 2008, ACM Press, 2008, pp. 110-115.

[12] M. Nagarajan, K. Verma, A. Sheth, and J. Miller, "Ontology driven data mediation in web services," Int. J. Web Service Res., 4(4), 2007, pp. 104–126.

[13] F. Saussure, "Course in General Linguistics", Open Court Publishing Co., Chicago, 1986.

[14] C. Thorpe and J. Bailey, "Commercial contracts, A practical guide to deals, contracts, agreements and promises", Woodhead, Cambridge England, 1996.

[15] UN Procurement Division, "Expression of Interest". http://www.un.org/Depts/ptd/eoi.htm.

[16] Walmart, "Request for Information". http://media.npr.org/assets/blogs/health/images/2011/11/Walmarthealthpartnerships.pdf.

 ISBN: 978-1-61208-280-6

Social Query: A Query Routing System for Twitter

Cleyton Souza, Jonathas Magalhães, Evandro Costa, Joseana Fechine
Laboratory of Artificial Intelligence - LIA
Federal University of Campina Grande
Campina Grande, Brazil
cleyton.caetano.souza@gmail.com, jonathas@copin.ufcg.edu.br, ebc.academico@gmail.com, joseana@dsc.ufcg.edu.br

***Abstract*—Social Query is a new and efficient way to get answers on the social networks. However, the popular method of sharing public questions could be optimized by directing the question to an expert, a process called query routing. In this work, we propose a Social Query System for query routing on Twitter, currently, one of the most popular social networks. The Social Query Systems analyzes the information about the questioner's followers and recommends the most suitable users to answer the questions. The use of the system changes the usual process, working apart of Twitter and allowing questioner and responder exceed the limit of 140 characters. Through a qualitative evaluation, we showed promising results and ideas for improving the system and the recommendation algorithm.**

Keywords- query routing; social query; expertise finding systems; community question and answering sites; social network; twitter

I. INTRODUCTION

Social query consists in sharing a problem (in the form of a question) with contacts in social networks and waiting for responses. It is an alternative to search engines and Community Question and Answering sites. Supported by popularity of social networks like Twitter [1] and Facebook [2], social query is a new and efficient way to find information on Web 2.0. The common strategy is to share a public question. However, this way is inefficient because after sharing a public question, there are several disappointing outcomes: receiving many answers (including wrong answers); keep receiving answers after having the problem solved; and never receiving an answer because people able to respond the question did not see it, since social network prioritizes visualization of most recent posts in Timeline, or did not feel an obligation to help [3, 4].

Horowitz and Kamvar [5] associate social query to the process of searching for answers in a village: when an individual in a village has a problem, before he goes to the library, he asks the most capable person that he knows; he does not ask everybody (the village paradigm). The same idea can be applied in the context of social networks. If the question is previously directed to someone (an expert), the social network will ensure that the expert will see the question through its notification system. However, choosing the right person is complicated and by choosing the wrong person, the author of the question may have to wait long for answers, may receive a wrong answer or may never receive an answer if the expert ignores the question [6]. The addition of Expertise Finding Systems to social networks would optimize the process and consequently enable quick and right answers [5]. The process of identifying an expert and directing questions to that expert is called query routing.

In this work, we present the Social Query System, a tool for query routing on Twitter. The system analyzes the information made available by the questioner's followers and ranks them based on three criteria: knowledge, trust and activity. Then, top users are recommended to the questioner who chooses to whom to direct the question. Details about the model [7] and the algorithm [6] could be obtained in our previous work. Our goal in this paper is to present how the system works and to show promising results of a qualitative evaluation performed with the first version of the Social Query System.

This remainder of this paper is organized as follows: Section II presents related work, Section III presents the usual process of sharing questions on the social networks, Section IV shows or proposal and how it works, Section V describes the results of our qualitative evaluation and Section VI presents our conclusions and future work.

II. RELATED WORK

Search engines are not always the best way to find information on the Web. Some problems are better solved by people (e.g., high contextualized questions, recommendations request, opinions request, advices request, and social connection request) [5]. An alternative to these types of problems is Community Question and Answering sites like Yahoo! Answers, which consists of online communities where users publish and answer questions voluntarily. After publishing a question, the user waits for answers from other users, who usually are unknown. However, people prefer to pose questions to their close friends in social networks rather than to unknown persons in Community Question and Answering sites [3, 5].

Regarding sharing questions on social networks, Morris et al. [8] presents statistics confirming that sharing questions is a viable method to obtain answers online. In their case study, 93.5% of users had their questions answered. In 90.1% of the cases, responses were provided within one day. However, that case study was conducted with Microsoft employees only, who possibly know each other and also use status messages from chat to ask questions. Paul et al. [4] conducted a similar study, but using only

 ISBN: 978-1-61208-280-6

Twitter users. They conclude that, in this specific context, only a only a small percentage of questions received answers (18.7%) and receiving an answer or not was strongly connected to the number of followers the questioner had. However, questions posted on Twitter are usually answered quickly. In their study, 67% of the responses come within the range of 30 minutes and 95% within the range of ten hours. These facts are due some features of Twitter: when a user posts questions to all followers, only a portion of these followers will view it and a smaller portion will respond. Thus, users with more followers are more likely to get answers, because there is a larger viewing of their messages. And, with respect to agility in receiving a response, this is mainly due to the nature of Twitter as a real-time social network. Actually, these statistics also show that, even as a regular strategy to obtain answers online, the social query process could be improved. We believe these results could be improved by applying query routing: after identifying an expert on the topic of the question and directing the question to him, the answer could come faster and with higher quality. Horowitz and Kamvar [5] establish a correlation between social query and the village paradigm: when people in a village are looking for information, before consulting the libraries, they first ask the most intelligent people they know.

In fact, the query routing problem could be understood as Expertise Finding problem. The query routing consists of a recommendation algorithm (or a technique) that finds an expert present in a group and directs a question to that expert [7]. In [9], it is presented a probabilistic and decentralized model for the question routing problem. This means that there is not an entity that makes all decisions and the routing algorithm works based on the repetition of actions taken in past. Probably, it was the first work about the query routing problem, but no system is proposed and the model is validated using simulated networks.

Davitz et al. [10] present a centralized model implemented in a tool called iLink. In this system, there is a global entity that monitors social network and decides who will receive questions, named super-node. Sometimes, the super-node is also able to offer answers. Other examples of systems are Aardvark [5], a social network that belonged to Google, and Q-Sabe [11], an academic tool for exchange of information focused on education. Both systems consist of Community Question and Answering sites where users could publish questions (questioners) that were routed to other users (responders) and these choose either answering or ignoring the question. Another example of query routing system is AskWho [12], a Facebook plugin that helps in the addition of mentions in the question. Silva et al. [13] proposed SWEETS, an Expertise Finding System, to AMIGOS, an academic online social network of Federal University of Pernambuco. Their recommender system monitors users and suggests experts based on a reading and writing profile.

The studies of Andrade et al. [11] and Horowitz and Kamvar [5] proposed query routing techniques and developed environments where they will work. Our research follows the reverse path. Our system works in a pre-existent context: Twitter, one of most popular social networks currently and that, apparently, will benefit of our technique [4]. In [12], it is presented a Facebook plugin, but AskWho does not use any special technique to match friends, consisting only in a search engine comparing keywords of the question with the profile of friends. Davitz et al. [10] and Silva et al. [13] also propose a system for pre-existent context, but iLink is only available to small communities due the computational effort to monitor the entire social network and SWEETS was not considered useful by AMIGOS users, being rejected by more than half of users [13]. Table I presents a comparison between the cited works.

TABLE I. WORKS COMPARISON.

Reference	Kind of Software	Recommendation Context	Features of Recommendation	Limitation	It uses an Activity criterion	It uses Relationship criterion
Andrade et al. [11]	Desktop.	Everyone who downloads the software.	Information Retrieval method.	Looking for people is not like looking for documents.	Yes (more active users are prioritized).	No.
Davitz et al. [10]	Web.	Users from Small Communities (e.g., forums, blogs).	Probabilistic method.	Only available to small communities.	No.	No.
Horowitz and Kamvar [5]	Web.	Users from Aardvark (a CQA site that belonged to Google)	Probabilistic method.	The project was closed in 2010.	Yes (more active users are prioritized).	Yes (users can optionally maintain a personal network).
Liu [12]	Web.	Friends on Facebook.	Keywords Matching.	Very simple.	No.	No.
Silva et al. [13]	Web.	AMIGOS Users.	Information Retrieval Method.	It was bad evaluated by the users.	No.	No.
Social Query System	Web.	Followers from Twitter.	Multi-criteria decision making method.	The current version only is available to Twitter users previously registered.	Yes (more active users are prioritized).	Yes (based on talks and similarities).

 ISBN: 978-1-61208-280-6

Our system uses three criteria during the recommendation process: knowledge (captured through the vocabulary used by followers), trust (captured through tweets exchanged and friends and followers in common) and activity (consisting in the mean latency time among messages). Besides, the followers are ranked using a multi-criteria decision making process called Weight Product Model (WPM). The use of multi-criteria is the main characteristic of the new generation of recommender systems and the WPM is the better method for the amount of criteria considered by us [14]. While most previous systems consider only the expertise of the candidates we use additional criteria related to the candidate's availability and the relationship between questioner and expert candidate. The next section describes the usual process of sharing questions on Twitter

III. THE USUAL WAY OF QUESTION AND ANSWERING ON TWITTER

Twitter is a microblog, a type of blog with some limitation of the content, where users can tweet (post a message) about any topic using 140 characters. In less than three years, Twitter reached such popularity that became the microblog with the largest number of users. Currently, five years after its creation, there are more than 200 million of users and daily registrations are about 460 thousand new users [15]. Another impressive data is the amount of posted messages: in January 2009 two million of tweets (messages) were sent per day; in January 2010 were 65 million and in 2011 were sent 200 million daily tweets [15]. According with the last numbers released by Twitter, currently, there are more than half billion of Twitter accounts and 340 millions of tweets being published per day [16]. This rapid growth has increased the interest of the scientific community on this online social network [1].

On Twitter, users can follow and be followed by other users. In this context, to follow a user means exposing interest in the content published by that user. The account of a user may be public or not, and in order to follow a protected account it is necessary to get the permission of its owner. Initially, the tweets (posts) are visible only to followers, they see in their timelines the tweets of those who they are following. Among the reasons that lead a user to follow another are admiration, friendship and reciprocity. In addition, a user may want to follow another one who posts content which may be considered relevant. Any user is allowed to reference others within a tweet and users can filter their mentions. Because of these features, many users use the microblog as a public chat [1].

When a user publishes a public question on Twitter, if it is not answered quickly, the chances of being visualized and answered in future decrease because the question will fall down in the followers' timeline. In Fig. 1, it is illustrated the process of publishing a question on Twitter.

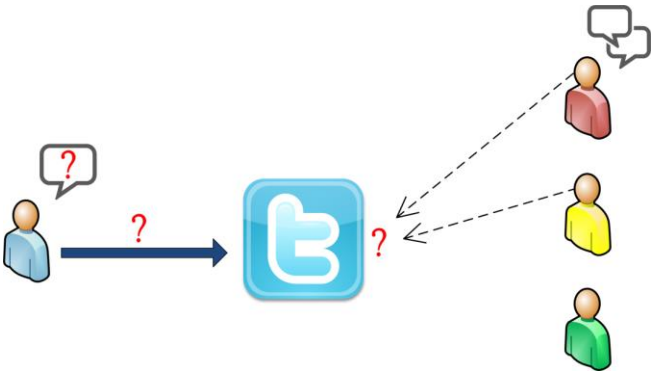


Figure 1. User Twitting a Question

After publishing a question, probably not all users who follow the questioner (the blue user) will see the question. Among the users that visualize (the yellow user and the red user), only a few will answer (the red user) and there is no guarantee that any of these answers will satisfy the questioner. Some users will not provide an answer because, as the question was posted for all followers, they do not feel an obligation to help. As the time passes by, the chances of the question being viewed and consequently answered in the future decreases, because it will fall positions on the timeline of the questioner's followers.

When a tweet (question) is previously directed to someone, the probability of it being visualized is much larger, because the mentioned user can filter the messages which mention him/her. In Fig. 2, it is illustrated the same process, but directing the question to a specific user (the green user).

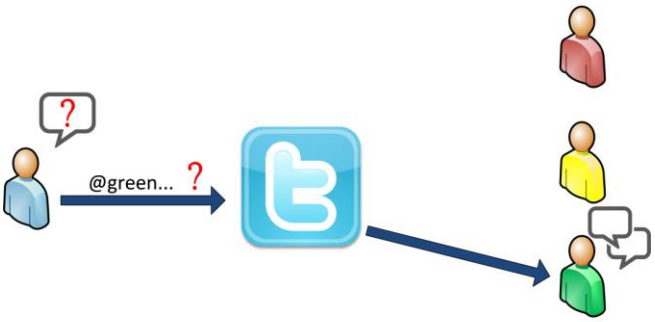


Figure 2. User Twitting a Direct Question to Someone

When a user (the blue user) mentions another user (the green user), the mentioned user immediately receives an email informing about the message. The mentioned user may wish to disable these notifications, but any user can filter their mentions, as already commented. Given these facts, we believe that direct the question to someone, in practice, guarantees that the message will be visualized by this person, but there is no guarantee that the message will be answered, neither about the quality of the response.

It seems evident that directing a question increases the probability of it being visualized, while the probability of receiving a good answer depends on whom it will be directed to. A query routing model consists in a recommendation

 ISBN: 978-1-61208-280-6

algorithm that examines the information available on the social network to decide who is able to respond the question. In Fig. 3, it is illustrated the query routing process working.

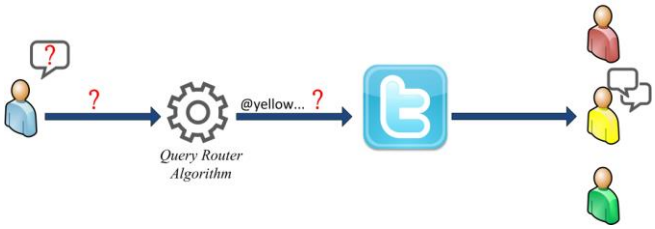


Figure 3. Using Query Routing to Tweet a Question

The question is formulated without a mention. The query routing algorithm (or routing algorithm) analyzes the information about the questioner's followers and ranks them, according to their ability to offer a right and quick answer. Then, the algorithm adds a mention in the question. In Fig. 3, the question is being directed to only one follower (the yellow user) but, as the algorithm ranks all candidates, it would be possible send the question to the *n* followers in best positions.

IV. THE SOCIAL QUERY SYSTEM

Our system works outside Twitter. Users access our system to ask and answer. Then, they tweet about their questions and answers. To exemplify how it works, we are considering two basic entities: the questioner (the author of the question) and the responder (a user recommended by our system and chosen by the questioner to receive the question).

Basically, in the query process, the questioner accesses the Social Query System and informs his Twitter account and the question (he also could add keywords to improve the recommendation process). The system analyzes information about the questioner's followers and recommends the top 5 users. Then, the questioner chooses to whom direct the question and tweets a message informing the chosen users (the responder) about the question. After clicking on the link informed in the tweet, the responder is directed to a page where the question can be answered, or someone else can be recommended to respond. Then, another message is tweeted informing about what was done. Working this way, the system has the advantage that questioners and responders could exceed the limit of 140 characters. Next, we will present in more details how the system works.

Fig. 4 presents the Homepage of our system. In this page, it is explained to the user how the system works and there is a link to the New Question page. While our system is on trial stage, it is only available for users previously registered by us. To require access it is needed to contact one of the authors. The Social Query System is available in two languages: English and Portuguese.

Figure 4. Homepage of the Social Query System

When the user clicks the "I HAVE A QUESTION" link, he is directed to the New Question page. This page is presented in Fig. 5. As already commented, there are three text fields: the questioner's account name on Twitter, the question and keywords (optionally).

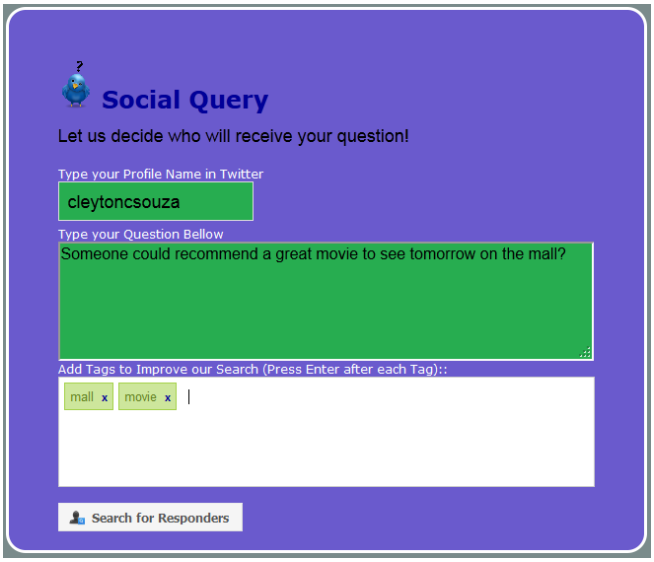


Figure 5. New Question Page of the Social Query System

When the questioner clicks the "Search for Responders" button, the system checks if the required text fields were filled and if the questioner is authorized to use the system. Then, if it is all right, the system analyzes information about the questioner's followers and ranks them based on three criteria: knowledge (an attribute that relates follower and question), trust (an attribute that relates follower and questioner) and activity (an attribute that relates followers and the social network). The top five candidates are recommended as responders on the Show Recommended List page, as presented on the left side of Fig. 6. If no user is a good recommendation, we suggest the questioner to reformulate the question or change the keywords.

 ISBN: 978-1-61208-280-6

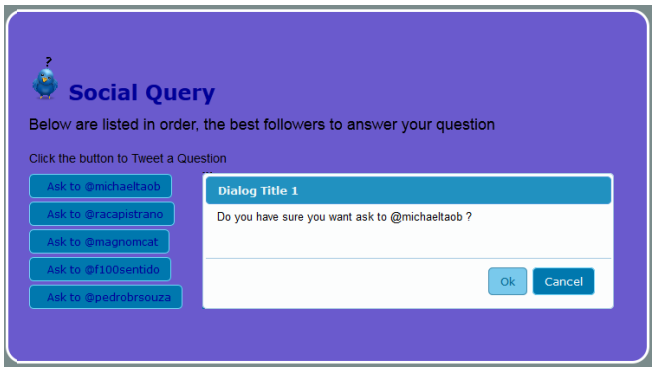


Figure 6. Show Recommendation List Page

Fig. 6 presents the recommendations. When the user clicks someone's name, a modal appears to confirm the questioner's intention to send the question to that person. When the questioner confirms, pressing the "OK" button, a new window is opened asking him to tweet the message "@questioner has a question for u @responder. Access http://short.link to answer. Thank u". The tweet has the following information: mention to the questioner, mention to whom was chosen by the questioner and a link to answer the question. After tweet the question, the questioner can repeat the process with the other users that were recommended or just close the page and wait for the answer. Someone who clicks the short link in the tweet will be directed to the New Answer page of the Social Query System, presented in Figure 7. In this page, the responder can give an answer (left button), inform that he/she does not know the answer (mid button) or recommend someone else to answer the question (right button).

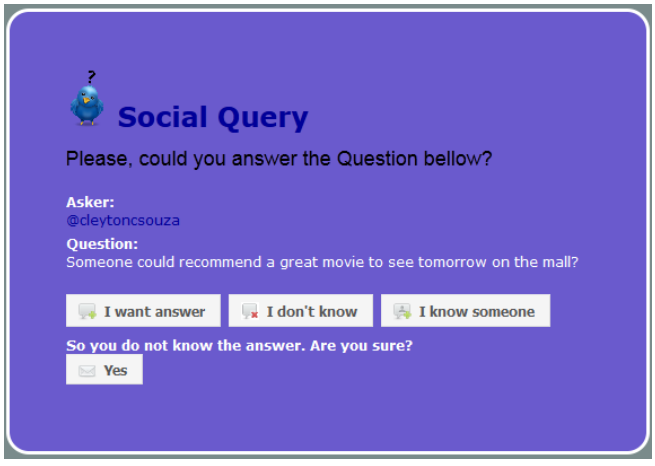


Figure 7. New Answer Page of the Social Query System

If the responder prefers to *suggest someone* else to answer the question, he clicks the "I know someone" button and informs his recommendation in a text field, and then tweets a message alerting the recommended user and the questioner about what this. The person recommended by the responder will click on a link in the tweet and will be directed to the New Answer Page. A responder who *does not know the answer* will press the "I don't know" button, and then will be directed to a page to confirm that he/she does not know the answer. The responder will be asked to tweet a message to the questioner informing about it. A responder who *knows the answer* clicks the "I Want Answer" button and is directed to the page presented in Fig. 8.

Figure 8. New Answer Page – Answering the Question

The responder writes the answer and clicks the "Send Answer" button. Then, a new window is open asking him/her to tweet the following message informing the questioner that the question was answered: "@questioner, @responder answered your question. Access http://short.link to see his answer. Thank u". After clicking the short link, the questioner is directed to the "New Evaluation" page, where the answer can be seen and have its quality evaluated. This page is presented in Fig. 9.

Figure 9. New Evaluation Page

The questioner clicks the "Evaluate" button and the evaluation is saved, ending the Question and Answering process. If the questioner sent the question to other responders, each responder will give an independent answer and the questioner will evaluate them individually.

V. EVALUATION

To validate our purpose, a qualitative evaluation was conducted. The goal of the evaluation is to analyze the opinion of volunteers about the recommendations made by the Social Query System. We used the following questions to perform the recommendations:

a) Looking for a new band to listen during weekend, does anyone have an indication?

b) Going to the movie theater after years LOL. What is the best movie in theaters?

Then, nine volunteers evaluate ten recommendations for each question. Each recommendation was labeled as *good*, *bad* or *neutral*. A recommendation was *good*, if the questioner believes that a relevant answer will be received. A recommendation was *bad*, if the questioner believes that an answer will not be received or an irrelevant answer will be received. A recommendation was *neutral*, if the questioner has no clue about the recommended person. For each question and each volunteer we calculate the percentile of good recommendations and compute the Normalized Discounted Cumulative Gain (nDCG), by considering *good* recommendation with a relevance of "1" and *neutral* and *bad* with "0".

In Table II, we present the main results of the evaluation.

TABLE II. SUMMARY OF RESULTS

Cases	Amount of Followers	% of good	%of bad	nDCG
Best case for Question "a"	192	50%	10%	0.63
Worst case for Question "a"	129	30%	60%	0.25
Best case for Question "b"	121	60%	0%	0.74
Worst case for Question "b"	68	30%	0%	0.18
Average for Question "a"	**110**	**41%**	**28%**	**0.41**
Average for Question "b"	**110**	**50%**	**20%**	**0.51**

According to Table II, the recommendations for Question "b" were better evaluated by volunteers. That happens due the way as the Question "b" was formulated: while Question "a" is composed by few words that express its subject (e.g., Band), Question "b" has more significant words and they are repeated (e.g., Movie and Theaters). This fact affects the knowledge model.

With respect to the worst cases, both users who classified most part of recommendation as *bad* or *neutral* informed that they are followed by many unknowns and they do not use Twitter to interact with friends but, mostly, to read and share news. With respect to the best cases they informed that they interact often with their followers and that follow back when they are real world friends. All these comments let us to add a new filter to recommendations related to the following back condition. Anyway, the mean performance of the Social Query System was considered very promising, almost half of recommendations were labeled as *good* and the *good* recommendations are better positioned in recommendation list.

VI. CONCLUSION AND FUTURE WORK

The main contribution of this work was the development of a tool to help people finding answers on social networks by improving the conventional social query process. This is a new and interesting research field with many limitations, mainly related to evaluation methods. Datasets provided by Community Question and Answering sites do not have all information available on social networks (e.g., relationship between users, messages directed to specific users, etc.) what hinders the use of a quantitative evaluation method. Qualitative methods have been used to evaluate recommender systems for a long time, but their results are very subjective and hard to compare. To evaluate our tool we used a qualitative evaluation method and we achieved promising results: more than half of recommendation list was considered useful by the volunteers. In addition, based on the opinion of volunteers, we added new features to the Social Query System, for instance, a filter to recommend only followers who are followed by the questioner and the inclusion of a new temporal and non-uniform criterion in recommendation that we call Availability.

As a future work, we are planning a quantitative evaluation of our model. We are creating a dataset and a quantitative method for evaluate our approach and compare with some previous work. In addition, another future work is a new qualitative study about real data collected with the mobile version of the Social Query System. Currently, we still projecting this app, but, probably, such application will be used by hundreds or thousands of users.

ACKNOWLEDGMENT

We want to thank all the volunteers who agreed to participate in this study.

REFERENCES

[1] B. Huberman, D. Romero and F. Wu, "Social networks that matter: Twitter under the microscope," First Monday, vol. 14, 2009, pp. 1-8.

[2] Y. Mui and P. Whoriskey, "Facebook passes Google as most popular site on the Internet, two measures show," The Washington Post, 2010.

[3] M. Morris, J. Teevan and K. Panovich, "Comparison of information seeking using search engines and social networks," Proc. 4th International AAAI Iternational Conference on Weblogs and Social Media (ICWSM), AAAI Press, 2010, pp. 291-294.

[4] S. Paul, L. Hong and E. Chi, "Is twitter a good place for asking questions? a characterization study," Proc. Fifth AAAI International

 ISBN: 978-1-61208-280-6

Conference on Weblogs and Social Media (ICWSM), 2011, pp. 578-581.

[5] D. Horowitz and S. Kamvar, "The anatomy of a large-scale social search engine," Proc. of the 19th International Conference on World Wide Web (WWW), ACM Press, 2010, pp. 431-440.

[6] C. Souza, J. Magalhães, E. Costa and J. Fechine, "Predicting potential responders in twitter: a query routing algorithm," Proc. 12th International Conference on Computational Science and Its Applications (ICCSA), Springer, 2012, pp. 714-729.

[7] C. Souza, J. Magalhães and E. Costa, "A formal model to the routing questions problem in the context of twitter," Proc. IADIS WWW/Internet (ICWI), IADIS Press, 2011, pp. 153-160.

[8] M. Morris, J. Teevan and K. Panovich, "What do people ask their social networks, and why?: a survey study of status message Q&A behavior," Proc. 28th International Conference on Human Factors in Computing Systems (CHI), ACM Press, 2010, pp. 1739-1748.

[9] A. Banerjee and S. Basu, "A social query model for decentralized search," Proc. 2nd Workshop on Social Network Mining and Analysiss, ACM Press, 2008.

[10] J. Davitz, J. Yu, S. Basu, D. Gutelius and A. Harris, "iLink: search and routing in social networks," Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2007, pp. 931-940.

[11] J. Andrade, J. Nardi, J. Pessoa and C. Menezes, "Qsabe-um ambiente inteligente para endereçamento de perguntas em uma comunidade virtual de esclarecimento," Proc. First Latin American Web Congress (LA-WEB), IEEE press, 2003, pp.1-7.

[12] C. Liu, "AskWho: finding potential answerers for status message questions in social networks," unpublished, 2010.

[13] E. Silva, R. Costa, L. Schmitz and S. Meira, "SWEETS: um sistema de recomendação de especialistas aplicado a uma plataforma de gestão de conhecimento a introdução," Revista de Informática Teórica e Aplicada (RITA), vol. 18, 2011, pp. 153-160.

[14] E. Triantaphyllou and S. Mann, "An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox," Decision Support Systems, vol. 5, 1989, pp. 303–312.

[15] Twitter Blog. "#numbers", 2011. Available at: http://blog.twitter.com/2011/03/numbers.html. [retieved: May, 2013]

[16] Twitter Blog. "Twitter turns six", 2012. Available at: http://blog.twitter.com/2012/03/twitter-turns-six.html. [retieved: May, 2013]

 ISBN: 978-1-61208-280-6

Using Chaos Theory to Investigate the Development Process of the Most Popular Blog

Kwoting Fang/National Yunlin University of Science & Technology
dept. of Information Management
Yunlin, Touliu, ROC
fangkt@yuntech.edu.tw

Shu Chuan Wang/Feng Chia University
Institute of Electronic Commerce
Taichung, ROC
scwang@mail.fcu.edu.tw

Abstract—**Advances in information and networking technology, in terms of online recreation, have continuously set an overwhelming pace for rapid growth in new applications of Internet-based activities in personal lives such as Blog. The main purpose of this study is to explore whether chaos exists in the blog development and discover any particular point in time worth further discussion by conducting quantitative analysis of chaos theory in the time dimension. Given with emergent events that cause divergence and uncontrollable behavior, the results revealed that the development behaviors of popular blogs undergo dynamic evolution over time and the future development is uncertain and unpredictable.**

Keywords- Blog; Chaos theory; Lyapunov exponents

I. INTRODUCTION

Blogs evolved from the online diary owned by Justin Hall in 1994. In 2003 Google bought blogger.com. In 2005, Bill Gates, Chairman of Microsoft, pointed out that blog was one of the important network applications, while Business Week featured blog on its cover as the theme of its 2005 Spring issue [1]. With the extensive use of blogs, more and more people became famous by writing articles or drawing comic strips on blogs. Popular blogs with huge numbers of daily visitors and responses often attract advertisers' attention and the bloggers may be invited to give public speeches, publish books or appear on product promotion events; fame and fortune soon ensue. However, in one of the commentaries entitled "The Blogging Iceberg" provided by Peruses Development suggested that the majority of blogs only attracted a small number of readers, and often became inactive rather quickly [2]. The company conducted research on the blog population development in eight blog service providers in 2008 and found that 30% of the blogs were active for only one day, and up to 66% of the blogs were not updated for over two months. The average active period was 126 days overall [3]. This data shows that it is not an easy task to keep a blog over a long period of time.

The existing theoretical models suggest that most of the studies on blog management tend to assume linear relationships among members interactions [4]. However, a closer observation found that blogs were an open platform, and some topics caused heated discussions through internal (bloggers, readers) and external (readers) interactions. These emergent events drew the blogs from stabilization to dissipation, which eventually returned to stabilization after a period of time until the next heated discussion began. The interactions in each phase also served as the feedback for the bloggers to post the next topic. It showed that blogs were a changing dynamic system [5], and there should be a theoretical framework different from linear relationships to explain the dynamic changes in blogs. Levy [6] proposed that Chaos Theory, derived from natural science, could be used to explain the unpredictable nature of social phenomena and to reveal the hidden pattern which changed over time. A number of scholars had also proposed the similar application of chaos theory in the enlightenment of social phenomena [7,8]. In addition, the existing literature [4,8] found that most research in blog behavior observation was confined to a fixed period, and only a few were longitudinal studies. If an observation is carried out within an extended period of time, the blog behavior changes over time can be tracked.

Therefore, the main purpose of this study is to explore whether chaos exists in the blog development and discover any particular point in time worth further discussion by conducting quantitative analysis of chaos theory in the time dimension. The results of this study should provide valuable insights for researchers who can improve techniques used in study for identifying exactly chaos situations. It can also be helpful for managers who can better understand how to bridge the gap between demand for Blog users and to make advertising more effective to fulfill Blog users in a competitive era.

The paper is structured as follows: Section II presents a literature review in terms of blog and chaos theory. Section III explains research model followed by Section IV, which uses Lyapunov exponent to identify the chaos exist or not. Finally, Section V concludes the paper and provides further discussions.

II. LITERATURE REVIEW

Justin Hall, a student at Swarthmore College, started his web-based diary in 1994 and was considered the earliest form of blog [9]. In July 1999 pitas.com released the first

 ISBN: 978-1-61208-280-6

free and user-friendly blog software "pitas", and launched blogger.com. It was a big deal when Google acquired blogger.com in 2003. Microsoft Chairman Bill Gates introduced MSN Space blog with MSN integrated in 2003. Since then, Blog has been rapidly developed and spread on the internet [10].

Several researches have been conducted on blogging in different areas since blogs began gaining popularity in 2003. Trammell and Keshelashvili [11] found that many popular blogs were diary-type and often contain illustrated texts. Zhang [12] attempted to predict the sustainability of blogs with chaos theory. The research period spanned from the first day of the blog till the end of year 2006. He concluded that when the Lyapunov exponent decreased and the system became less chaotic, the bloggers should take the initiative to post new entries to improve the situation. Hsu and Lin [13] undertook research on acceptance of blog usage and found that perceived ease of use, entertainment, altruism, reputation, and community identification were important factors that influence the behavioral intention to adopt blog service.

Chaos Theory, originated from Mathematics and Physics in the 1960 and 1970 [6], argued that the direct causal relationship proposed by the Newton theory in which it believed "one cause produced one result" was only a special case in the real world. Chaos theory proposed that nonlinear system was widespread in nature, included biological growth and fall of ethnic groups. In a dynamic system, objects were extremely simple in operation at the beginning, and unexpected results were produced with continuous duplications of the status in the previous phases by a certain rule over time [14].

Concomitant with the markedly increasing number of Blog users, it is not surprising that Blogs have garnered significant attention from the academic setting at large, from scholars in disciplines as diverse as sociology, psychology, education, and information systems to explore the potential research issues.

In the field of finance, some scholars [15,16] used chaotic phenomena in well-log time series to analyze the changes in the United States Treasury bill rate and stock market price index. Their results confirmed that chaos theory was valid for explaining the real market situation. In information system based research, McBride [17] applied the concepts of initial condition, strange attractor and dissipative structure to interpret the interactions between the system and its members. Utilizing secondary data to review the revolution in monitor technology, Tu and Hung [18] applied chaos theory to explore the changing process of technical environment dynamic cycle.

III. RESEARCH METHOD

This study presents the research model (Figure 1) that analyzes of blog data and that is used to identify whether chaos exists.

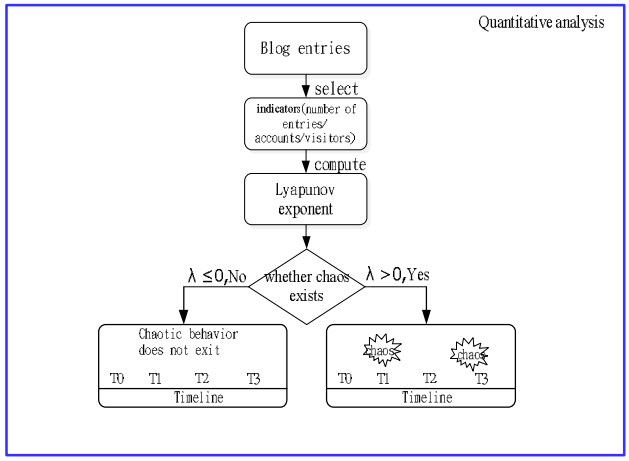


Figure 1. Research model

To investigate whether chaos exists in the most popular web blog, and to understand the driving events that cause chaos and its implications, we define "the most popular web blog" as "the blog which has the most visitors since the first day the blog". The data is obtained from web site [19] which began tracking blog popularity and ranking blog growth in real time since September 2006 (Blog Observation System, 2009). As of June 2009 there were 211,574 blogs on file. This study picked the top ranked Chinese personal blog *Wan Wan's Comic Blog* as the research object (Blog Observation Chinese Personal Blog, 2009).

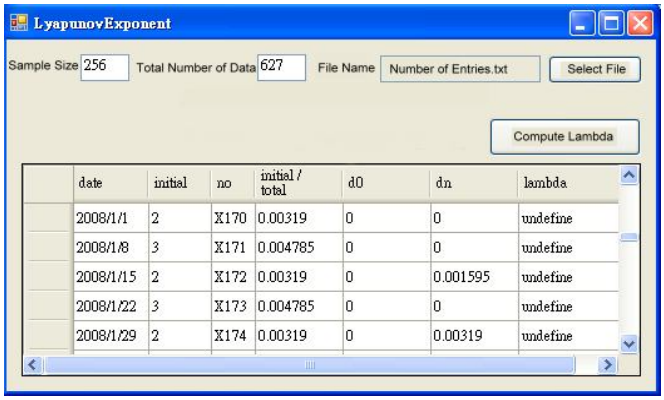
Identify suitable indicators for measuring the behavior of blog development, this study collected the information which is publicly available on the blog service website, including the number of entries, accounts and visitors as the indicators for measuring whether chaos exists during the course of the blog development, and analyzes the results to get the most discriminative indicator. For minimum sample size consideration, this study collects the data from the first day of the blog to the end of August 2009, and selects "week" as the appropriate time interval for the indicator analyses to meet the minimum sample size 200 required by Lyapunov exponent analysis [17]. *Wan Wan's Comic Blog* began in October 2004. Totally, 256 sample data are collected for each indicator (October 2004 to August 2009, a total of 64 months. 64 * 4 = 256 samples)

IV. DATA ANALYSIS

In line with above sense, the data was analyzed in two phases. The first phase calculated the total number of the entries, account numbers and visitors within one week. The number was used as λ in phase two for the Lyapunov exponent calculation to determine whether each indicator denotes chaotic behavior. The raw data was collected from the entries on Wan Wan's Comic Blog. The results of the Lyapunov exponent were shown in Fig.2 to Fig.4. Fig.2 was the λ value for entry number. To meet the minimum

 ISBN: 978-1-61208-280-6

requirement of sample number, the data was aggregated using one week as interval to minimize the difference in the number of entries for each week, which made the do or dn in the Lyapunov exponent equation λ = LN(dn/do) to be 0 and λ to be undefined. Therefore it could not be used as an indicator of chaos. Fig.3 and Fig.4 showed the trend map using account numbers and visitors for further analysis.



	date	initial	no	initial / total	d0	dn	lambda
	2008/1/1	2	X170	0.00319	0	0	undefine
	2008/1/8	3	X171	0.004785	0	0	undefine
	2008/1/15	2	X172	0.00319	0	0.001595	undefine
	2008/1/22	3	X173	0.004785	0	0	undefine
	2008/1/29	2	X174	0.00319	0	0.00319	undefine

Figure 2. λ value for entry number

Kiel & Elliott [7] proposed in the chaos theory study that for any system if there was at least one positive value ofλ, the system would have chaotic behavior, and the higher the value the greater the degree of chaos. The results from the Lyapunov exponent find that for both the account numbers and visitors, the λ values were greater than 0 during the period 2004/12/23 to 2009/4/21 (Fig.3 and Fig.4). Since there was at least one positive λ value, it could be assumed that chaotic behavior existed in the system. The result is shown in Table 2.

The analysis of Lyapunov exponent in Table 2 showed that the results of testing chaos using account numbers and visitors are quite consistent. During the period 2004/10/5 – 2004/12/22 the λ values were less than 0, which meant the blog was stable without chaotic behavior. However, during the period 2004/12/23 – 2009/4/21 there was at least one λ value greater than 0, which meant chaos existed in the blog development and the development had become uncertain and unpredictable.

According to Hilborn [20] the subsystems were formed with different initial values, we could further understand the extent of chaos in certain intervals by the $\overline{\lambda}$ values. To better observe the chaotic behavior in the system the study uses "one year" as the time interval and observes the changes of the $\overline{\lambda}$ values in each interval to understand the changes in chaos each year. The reason for selecting one year as the interval was to avoid the low number of entries and critical events in a short period of time. Table 3 and Fig.5 showed the λ values of account numbers and visitors in each year.

Table 3 used account numbers as an indicator for chaos and showed that the $\overline{\lambda}$ value before 2004/12/22 was less than 0 ($\overline{\lambda}$ =-0.440); this meant chaotic behavior did not exist during the development. However, the values moved from negative to positive in 2005-2009 and went up increasingly, which meant that chaos existed in the blog development and the extent of chaos was becoming more significant. The value was highest in 2008, indicating there was a critical event worth attention during this period. However, the value in 2009 decreased slightly compared to 2008. It might be attributed to the requirements for Lyapunov exponent calculation. The data collecting stopped at the end of April 2009, which affected the $\overline{\lambda}$ value for the year 2009. The analysis result of the $\overline{\lambda}$ value for visitors was consistent with that for account numbers.

Table 2 and Table 3 both revealed that the identification of chaos using account numbers and visitors is consistent with both one year or the entire period as an interval, which meant both indicators were valid in identifying chaos. The change in $\overline{\lambda}$ value each year showed that blog development involved both positive and negative forces. The negative force led the system to a stable status, and with certain events the system diverged from stability. Table 3 showed the $\overline{\lambda}$ value changed from negative to positive at the end of 2004 to 2005. The positive force led the system away from stability to chaotic behavior.

In this section we adopted the chaotic behavior - time matrix to express chaos in blog development. (Fig.6)

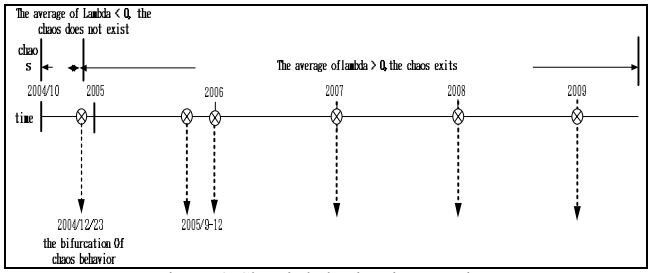


Figure 6. Chaotic behavior-time matrix

The Lyapunov exponent analysis showed that both positive and negative feedback forces existed in the blog development process. The positive feedback force meant the system was chaotic, divergent, and was unpredictable and difficult to explain. The negative feedback force meant the system was stable and seeking internal development and recognition. The system was relatively closed. In the initial stage the blog was not well-known and only attracted a few visitors, therefore the $\overline{\lambda}$ value prior to 2004/12/22 was smaller than zero, meaning chaos did not exist and the blog was stable and convergent. The system underwent a

 ISBN: 978-1-61208-280-6

transition from non-chaotic to chaotic on 2004/12/23. The λ values from 2005 through 2009 were greater than zero and were steadily increasing; indicating critical events existed during these periods and led the system to go from the state of equilibrium to disorder.

V. CONCLUSIONS AND DISCUSSIONS

The main purpose of the study is to explore whether chaos exists during blog development and find out the timing when this behavior occurs. The results show that the non-linear and dynamic chaos does exist during the course of the blog development and there exists both positive and negative forces in the system. Basically, the negative force leads the blog towards stability and convergence, and during this period the blog seeks readers' recognition and cohesion. Critical events bring in the external energy and resources, which causes internal and external interactions under the dissipative structure of the system. In contrast, the positive force therefore leads the system to move from the state of stability towards disturbance, divergence and disorder and makes the situation difficult to predict and control. The time, chaos occurs, is when Lyapunvo exponent turns from negative to positive. This is the demarcation of ordered structure and disorder, meaning there're emergent events that cause divergence and uncontrollable behavior. These are events that require discussion to understand the applications of the quantitative data. The demarcation of the negative force and positive force is on December 23rd 2004 when MSN emoticons are released. The blog goes into a state of chaos and unpredictable situation. In addition, the $\overline{\lambda}$ values for accounting numbers/visitors (Fig.4) from the result of quantitative data analysis show that the $\overline{\lambda}$ values have increased over the years since the demarcation and reached the highest number in 2008, meaning the degree of chaos disturbance peaks in 2008

The chaotic behavior - time matrix (Figure 6) shows that under the operation of self-organization feedback mechanism, the blog develops diversely and grows continuously. Although the blog features the blogger's personal style and taste, it needs to be further extended and broadened to attract more visitors. From a commercial viewpoint, specific topics are needed to generate discussion, however in addition to the linear thinking where cause and effect have a direct relationship, it should be noted that small cause may also bring large effects. Finally, the virtual online community is not sufficient, face-to-face interactions also help boost popularity. The update on the blog finds heated discussion after the Wan Wan's first face-to-face interaction with the fans.

ACKNOWLEDGMENT

The authors would like to thank the National Science Council of Taiwan for partial financially supporting this research under NSC 100-2410-H-224-002-MY2.

REFERENCES

[1] Business Wire. The blogging iceberg: Of 4.12 million weblogs, most little seen and quickly, abandoned, according to perseus survey. Retrieved on May, 31, 2013, from the: http://www.businesswire.com/portal/site/home/

[2] R. Harmanci. Time to get a life - pioneer blogger Justin Hall bows out at 31. Retrieved on May, 31, 2013, from: http://www.sfgate.com/cgi-bin/article.cgi?file=/c/a/2005/02/20/MNGBKBEJO01.DT

[3] T. A. Brown. Measuring chaos using the lyapunov exponent. In L. K. Douglas & E. Elliott (Eds.). *Chaos theory in the social science, pp*.53-66. Ann Arbor, MI: University of Michigan Press, 1996..

[4] P. V. Mcdonough, J. P. Noonan, & G. R. Hall. A new chaos detector. *Computers and Electrical Engineering*, vol. 21(6), pp. 417-431, 1995.

[5] S. H. Kellert. In the wake of chaos: Unpredictable order in dynamical systems. USA, Chicago: University of Chicago Press, 1993.

[6] D. Levy, D. Chaos theory and strategy: Theory, application, and managerial implications. *Strategic Management Journal*, vol. 15(2), pp. 167-178, 1994.

[7] L. D. Kiel, & E. W. Elliott. Chaos theory in the social sciences: Foundations and applications. Ann Arbor, MI: University of Michigan Press, 1997.

[8] L. L. Hsu, V. Xu, & H. C. Wu. The antecedents of influencing usage intention in blog context. *Journal of E-Business*, vol. 11(1), pp. 1-28, 2009.

[9] J. Gleick. Chaos: Making a new science. New York, NY: Penguin, 1987.

[10] Blog Observation System. Retrieved on May,31, 2013, from : http://look.urs.tw/hitsrank.php?type=1

[11] K. D. Trammell,, & A. Keshelashvili. Examining the new influencers: A self-presentation study of a-list blogs. *Journalism & Mass Communication Quarterly*, vol. 82 (4), pp. 968 – 982, 2005.

[12] J. X. Zhang. A study on the prediction of blog duration. Unpublished master's thesis, National Chung Cheng University, Chiayi, 2006.

[13] C. L. Hsu, & J. C. Lin. Acceptance of blog usage: The roles of technology acceptance, social influence and knowledge sharing motivation. *Information and Management*, vol. 45(1), pp. 65-74, 2008.

[14] S. Irene. Chaos theory and institutional economics: metaphor or model? *Journal of Economic Issue*, vol. 33(1), pp. 141-167, 1999.

[15] R. F. Bobner, I. Newman, & C. Wessinger. Chaos modeling: Increasing educational researchers' awareness of a new tool. *Proceeding of the Mid-Wester Educational Research Association*, Chicago, IL, 1989.

[16] T. Miyazoe, & T. Anderson. Learning outcomes and students' perceptions of online writing: Simultaneous implementation of a forum, blog, and wiki in an EFL blended learning setting. *System*, vol. 38(2), pp. 185-199, 2010.

[17] N. McBride. Chaos theory as a model for interpreting information systems in organization. *Information Systems Journal*, vol. 15(3), pp. 233-254, 2005.

[18] M. F. Tu, & S. C. Hung. Finding cycles in technological change: An analysis of displays from 1976-2003. *Journal of Management*, vol. 25(3), pp. 291-308, 2008.

 ISBN: 978-1-61208-280-6

[19] The data is obtained retrieved on September 20 2012 from http://look.urs.tw/

[20] R. C. Hilborn. Chaos and nonlinear dynamics. New York, NY: Oxford University Press, 1994.

TABLE 1 Data format

Number of sample	Entry date	week	Lyapunov exponent analysis indicator		
			number of entries	number of accounts (responses +trackbacks)	visitors
1	October 2004	Week1	data	data	data
⋮	⋮	⋮	⋮	⋮	⋮
256	August 2009	Week4	data	data	data

TABLE 2 λ value for account number/visitors

Exponent	2004/10/5~2004/12/22	2004/12/23~2009/4/21
Account Numbers	The system is stable (all λ values in the interval are <0)	The system is chaotic (at least one λ value>0 in the interval)
Visitors	The system is stable (all λ value in the interval are <0)	The system is chaotic (at least one λ value>0 in the interval)

TABLE 3 $\overline{\lambda}$ value for account numbers/visitors

indicator	Prior to 2004/12/22	2004/12/23	2005	2006	2007	2008	2009
Account numbers $\overline{\lambda}$	-0.440	0.403	0.0049	0.1901	0.3416	0.4023	0.3137
Visitors $\overline{\lambda}$	-0.389	0.427	0.0170	0.0588	0.3121	0.4138	0.3331

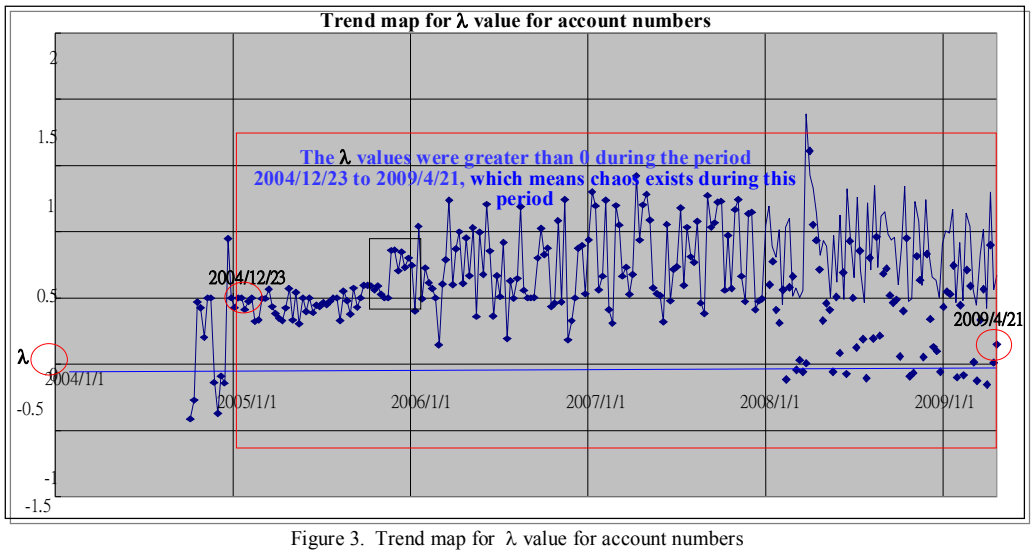


Figure 3. Trend map for λ value for account numbers

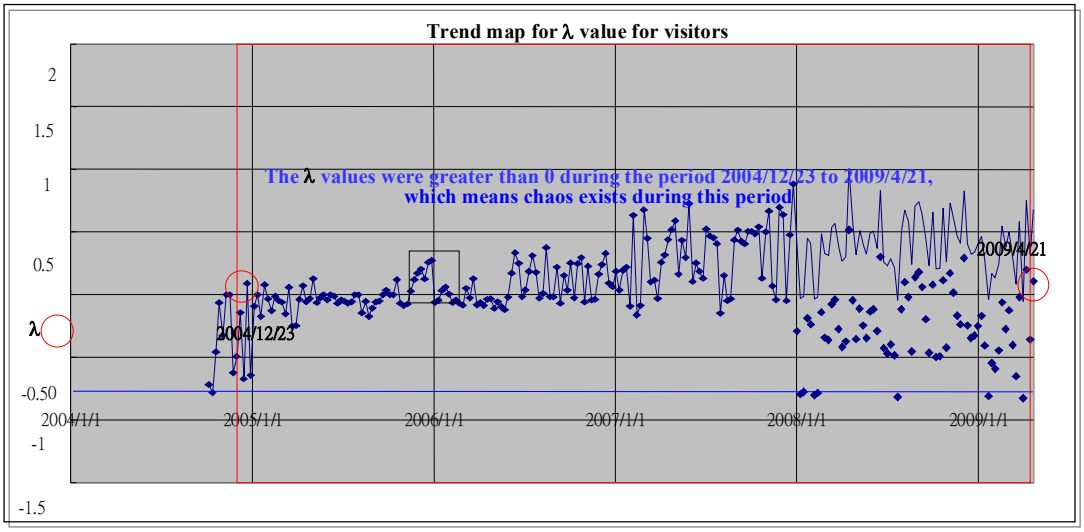


Figure 4. Trend map for λ value for visitors

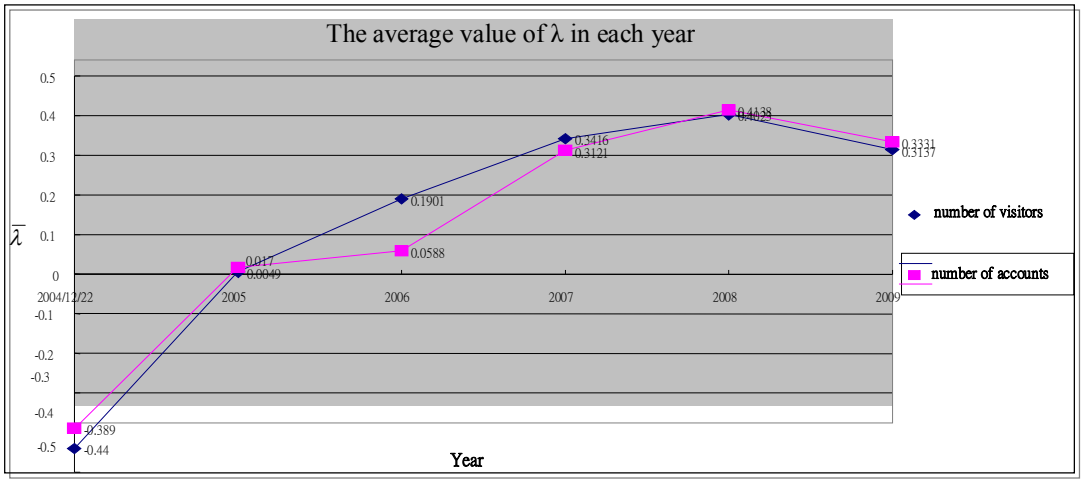


Figure 5. $\overline{\lambda}$ Value for account numbers/visitors

Stigmergy within Web Modelling Languages: Positive Feedback Mechanisms

Aiden Dipple, Kerry Raymond, Michael Docherty
Faculty of Science and Technology
Queensland University of Technology
Brisbane, Australia
aiden.dipple@student.qut.edu.au
k.raymond@qut.edu.au
m.docherty@qut.edu.au

Abstract— **Stigmergy is a biological term originally used when discussing insect or swarm behaviour, and describes a model supporting environment-based communication separating artefacts from agents. This phenomenon is demonstrated in the behavior of ants and their food foraging supported by pheromone trails, or similarly termites and their termite nest building process. What is interesting with this mechanism is that highly organized societies are formed without an apparent central management function. We see design features in Web sites that mimic stigmergic mechanisms as part of the User Interface and we have created generalizations of these patterns. Software development and Web site development techniques have evolved significantly over the past 20 years. Recent progress in this area proposes languages to model web applications to facilitate the nuances specific to these developments. These modeling languages provide a suitable framework for building reusable components encapsulating our design patterns of stigmergy. We hypothesize that incorporating stigmergy as a separate feature of a site's primary function will ultimately lead to enhanced user coordination.**

Keywords-web collaboration; virtual pheromones; stigmergy.

I. INTRODUCTION

The World Wide Web has transitioned from its historically static content to a new, dynamic experience emerging through collaborative websites and social networking. We seek to understand a specific set of emerging designs that we believe are indicative of a natural phenomenon called stigmergy [1].

In biology, stigmergy describes a mechanism of indirect communication where the actions of individuals affect the behavior of others (and their own). This communication mechanism describes what has been considered as apparent cooperative behavior of insects' during various activities. An example of this is the food gathering activities of ants which exploit pheromone trails. To find the most recent and relevant food source the ants select paths to follow based on the strength of specific trails. The environment embedded pheromones is considered a form of indirect communication. This stigmergic communication comprises of an explicit message in the pheromone to gather food, and an implicit signal through the current level of decay: information within the trails themselves show which trail will currently lead to a food source opposed to those leading to a depleted food source.

There are multiple varieties of stigmergy and our research to date has modeled and documented this [2, 3]. We have created generic design proto-patterns from observing stigmergy in numerous Web sites, however this paper will focus on those observed within Facebook. The User Interface (UI) designs observed provide representations of user feedback along with representations of behavior trends from unintentional interactions recorded as trace data. Examples of these UI designs can be seen in Facebook where users "Like" other user contributions causing an area of focused interest. Another example can be seen where Facebook has introduced a "Seen By" representation of feedback where the number of users navigating to an article is presented as a trail (or virtual footsteps).

Web Modeling Language (WebML)[4] is a method of modeling data content, user interaction and navigation flow for Web 2.0 applications. WebML provides a way to design the mapping of a data model to different UI views and the navigation paths between those views. The pertinent aspect of the WebML framework to our research question is that WebML is designed to be extensible to facilitate new concepts, interface types and event types. Our research uses WebML and the WebRatio development environment allowing stigmergy to be easily incorporated into a site as reusable components keeping the core code-base separate from the stigmergic features. Given the Web 2.0 UI designs that we have observed and a thorough analysis of how they correlate to stigmergy, we have implemented generic UI components as standard elements for web site development to exploit stigmergic communication.

Our research project focuses on creating a model of stigmergy that we can use to design feedback mechanisms into Web applications. From this model we can describe a

 ISBN: 978-1-61208-280-6

framework to exploit stigmergy in the collaborative environment provided by specific Web applications. This paper illustrates how we create design patterns in an Integrated Development Environment (IDE) to capture the features of stigmergy as a separate aspect to core Web site features. The research contribution of the PhD project is to determine the efficacy of incorporating stigmergy into site design; this paper makes a significant contribution by documenting our current progress capturing user input (positive feedback) using reusable stigmergic design patterns readily included into primary Web site functionality. Future work will involve testing our model of stigmergy using our design patterns in experiments with users to test how they react to the presence of stigmergic features in a Web 2.0 site when compared with a similar Web 2.0 site without the stigmergic features.

This paper will introduce our research question based on our hypothesis in Section II. Section III will provide a brief Literature Review describing previous work on Stigmergy providing key facets of our model development. Section IV will detail our research methodology and explain why we have chosen a multi-method approach. Section V will detail our current progress including an overview of our developed model of stigmergy, data model and interface components implementing the positive feedback mechanisms. Section VI will discuss problems yet to be solved and will be followed by the conclusion in Section VII.

II. RESEARCH PROBLEM

Our hypothesis is that stigmergic behaviour is inherent in collaborative Web environments and that a framework to support all attributes and dynamics of stigmergy will facilitate higher quality collaborative outcomes. This leads to the question: Does the Web enable us to build better collaborative sites when the attributes and dynamics of stigmergy are fully exploited? Are there facets of stigmergy missing in the Web environment that could be used in capturing implicit communication otherwise lost? To answer this question the project has required a clear definition of stigmergy and how it manifests in Web environments. There is significant research into stigmergy, virtual pheromones and swarm intelligence re-creating stigmergic behaviour; but limited research into its relevance as a design pattern to coordinate human behaviour. If we can build a model identifying stigmergy in Web environments, we speculate we can create a methodology to build sites benefiting from this phenomenon.

III. STATE OF THE ART

The word stigmergy "is formed from the Greek words stigma 'sign' and ergon 'action'" [5] and is used within biology to describe the way non-rational, autonomous agents (such as termites or ants) collaborate to achieve complex tasks thereby displaying some type of emergent swarm-intelligence [6]. These agents use pheromones as signs embedded within the environment to trigger behaviour or actions in other agents in the swarm.

Stigmergy was first introduced in 1959 by a French zoologist named Pierre-Paul Grasse [7] to describe how insects appear to coordinate successfully despite having no centralized management structure or direct observable intercommunication [1]. A simplified definition of stigmergy is: a process by which agents communicate indirectly between one and other through their environment. More specifically, the behaviour of agents is influenced or determined by the behaviour of agents which have interacted with the spatio-temporal environment previously [8]. Essentially stigmergy describes an autonomous system enabling self-organisation, self-optimisation and self-contextualisation in a light-weight and scalable mechanism [9]. This is interpreted as the associated mechanisms and emergent behaviour enabling the selection of the optimal solution without the prerequisite of knowing anything about the environment.

Stigmergy is a compelling phenomenon because it describes a positive feedback system where the signal strength of a trail will increase as more agents follow that trail. This leads to more rapid successful task completion. In opposition to this the environment enacts upon the sign causing atrophy and entropy to diminish the signal strength. This decay provides the negative feedback ensuring only the most current trails can be sensed and that old trails do not interfere with the task as they become redundant. Stigmergy provides a model of both *active contributions* and *passive interaction* with both varieties being demonstrated within the Web. The two varieties of stigmergy have been categorized as *marker-based* [1] and *sematectonic* [10]. Marker-based stigmergy describes an explicit modification of the environment by leaving a sign with the intention of signaling to other agents. Marker-Based stigmergy is broken into two sub-types: *qualitative* and *quantitative* [1]. The sub-types differentiate where single contributions are sufficient to elicit a response versus an accumulation of contributions increasing the probability of triggering a response as signal strength increases.

In contrast to explicitly leaving contributions, sematectonic stigmergy is defined as a modification to the environment as a by-product of actions being performed. These by-products are occurring inadvertently and unintentionally to the primary task being performed. For example, when considering a path being left in a lawn when people take a short-cut across it they have no intention of signaling to others that they have taken a short-cut. The short-cut is the purpose of the action, but the environment will retain the footstep impact as an alteration of the environment. There is no explicit foot-step left in the environment (obviously excluding cases such as wet feet leaving wet foot prints) however the action has altered the environment and the cumulative foot-step action manifests in the format of a path rather than something recognizable as an aggregation of individual feet traces. These two

 ISBN: 978-1-61208-280-6

varieties of stigmergy highlight the notion of intentionality of communication as being either explicit or implicit [11, 12], with marker-based stigmergy being explicit communication and sematectonic stigmergy being implicit communication.

There has been a significant amount of research focused on stigmergy in robotics and Web environments [5, 13]. Web environments provide a close facsimile to stigmergy in physical environments where a large number of users coordinate in a highly organized manner indirectly communicating through the contributions they make within the Web sites. Our research focuses on how the varieties of stigmergy manifest as Web environment User Interface (UI) elements and how they can be employed within Web site design to improve user collaboration and information categorization. We seek to do this using emerging web modeling technologies.

Web Modelling Language (WebML [4]) is a platform independent way to express the interaction design, data model and business rules of Web application development separately from the implementation platform. WebML permits the formal specification of the data model, interface composition and navigation options. WebML describes a visual notation for designing Web applications that to be exploited by the visual design tool WebRatio for the auto-generation of code. We have implemented proof-of-concept UI mechanisms to record and display both intentional and unintentional web site interaction based on our model of stigmergy that we present in this paper.

IV. RESEARCH METHODOLOGY

This research project focuses on identifying the attributes and dynamics of stigmergic behaviour and how it facilitates and benefits the process of recording *active contributions* and *passive interaction* of users when participating in the *grand purpose*. The research is based on a mix-method approach comprised of a literature review content analysis, comparative case studies of existing Web sites, and finally experimentation and data analysis testing the stigmergy design patterns created as part this research.

The literature review has provided a thorough analysis of stigmergy exposing the complexities of the phenomenon and how to best incorporate the properties of stigmergy into a Web environment. The results of the initial analysis stage has led to the development of a model describing the attributes and dynamics of stigmergy along with documentation tracing its components back to the work performed by previous researchers. This provides the chain of evidence to validate the model and enable its correctness to be reviewed.

Due to the qualitative nature of the data collection, a comparative case study approach has been used [2] to provide legitimacy to the repeatability of the research findings. The pattern in the developed model has allowed the comparative case study to be performed against a selection of existing web sites with varying levels of model alignment. Analysis of the case studies identified common solution patterns as well as *proto-patterns* representing solutions which provide more sophisticated implementation of the stigmergy varieties [3]. Targeting multiple sites for case studies has provided a vital cross section of sites displaying aspects that impact the simplistic entomological examples of stigmergy when applied to complex and cognitive human systems. Targeting multiple sites over a broad spectrum of social aspects of the Web has exposed the repetition of stigmergic patterns further enforcing the generic design of our model.

V. PROGRESS TO DATE

The investigative stages of the research plan have been completed including the literature review and initial case study. The literature review includes the analysis of stigmergy as a generic phenomenon and from the perspective of various algorithm designs. Previously [2], we have introduced the resulting model (see Figure 1) of stigmergy including the concept of a stigmergy *grand purpose* and the core components of stigmergy: the *agent*, the *environment*, and the *sign*.

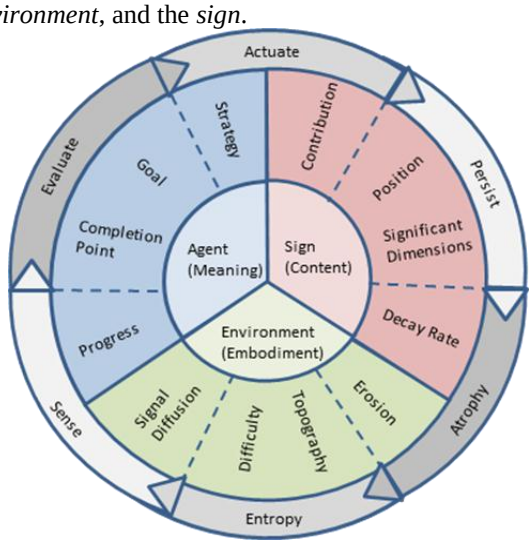


Figure 1. Stigmergy Cycle

Stigmergy facilitates a *grand purpose* (or emergent behaviour) through the *dynamics* (or mechanisms) applied to its inherent *attributes* (or components) of the environment, agents, and artefacts. Our progress to date has provided clarification and the categorisation of virtual pheromones (and other traces) and their role as triggers. The development of our model has provided insight into various similar indirect methods of communication that are considered to be a superset of stigmergy: Behavioural Implicit Communication (BIC). BIC is considered outside of our research area of stigmergic communication. Our modelling of the concepts of implicit and indirect communication mechanisms has provided the missing holistic, conceptual synthesis of the phenomenon. Our intended contribution of analysing the efficacy of stigmergy within Web 2.0 sites will build on this contribution.

 ISBN: 978-1-61208-280-6

The model ties together the core components of stigmergy: an inner band representing the attributes of the components; and an outer band representing the dynamics acting on those attributes. The outer band dynamics are either internal to each component, or defining the interface between components. The model describes the dynamics of equilibrium between positive feedback (contributions) and negative feedback (decay) illustrating how the agent contributes the positive feedback, where the environment applies the negative feedback. This is a generic model of stigmergy that applies for the world of entomology, the human world and the virtual world. This paper focuses on how the varieties of stigmergy manifest as Web environment User Interface (UI) elements; therefore the three components of stigmergy correlate to the users of Web environments, and the contributions that the users make.

To incorporate stigmergy into WebML we need to understand the ways the system receives input and how it should display output. The *contribution* attribute of the model correlates with the user having an input mechanism to *actuate* the contribution to a sign; the *signal diffusion* attribute defines accessibility of an output mechanism presenting the transformed contributions as a signal to users. The environment requires the capacity to store contributions therefore requiring a stigmergy specific data model.

Our research has explored how the different varieties of stigmergy manifest as proto-patterns within the target case study sites [3]. The key concepts identify that there are consistencies between the input (*actuators*) and output (*sensor*) mechanisms for each variety of stigmergy. These findings highlight that the observed Web site cases can be implemented using reusable stigmergic mechanisms.

The two simplest forms of input mechanism for marker-based stigmergy both enable a user to intentionally make a selection, whether as a single presented choice or as a single choice from a number of options. An example of the single presented option can be seen in Facebook with the "Like" feature where there is only one option presented to the user. An example of a single selection of multiple options can be seen in rating systems such as the one-to-five scale within eBay. More accurately the eBay example is a composite set of options where a group of categories are presented (e.g., communication quality, postage costs, etc.) with each choice selection aggregating into a single seller reputation metric. In the case of sematectonic stigmergy, the trigger for the contribution is hidden from the user and occurs unintentionally when the user interacts with site content or navigates to particular pages. An example of this is seen in the Facebook "Seen By" feature that records which users view a particular Group's news-feed item.

The two simplest output mechanisms observed correspond to signal type: quantitative or qualitative. The Facebook quantitative signal type illustrates how the contributions are transformation into an aggregate summation presented to the user as a "Liked" count. The eBay example provides a metric that is based on a more complex function but presented as a single percentage value. Our design facilitates both the storage of each of these types of contributions and each type of presentation.

The qualitative signal type is a detailed list of raw contributions and can be seen within the Facebook "Share" feature. The user contribution broadcasts specific content displayed within the standard Facebook news-feed. The contribution is a reference to an article and is stored as a primary key value and specific data model entity name that that key relates to. The actuator input mechanism to record the contribution is the same as for the Facebook "Like" example. The sensor output mechanism of this example is the propagation of the sign to the recipients with accessibility to the signal as defined by *signal diffusion*.

Our most recent progress has been to implement proof-of-concept examples within the WebRatio development environment built on these -generic proto-patterns. The data model for our tests can be considered within three separate components: core entities for application functionality; supporting entities (e.g., user accessibility entities); and stigmergy entities.

Figure 2 shows the user accessibility entities (*user*, *group* and *module*) that are created by default within WebRatio. Also shown is the stigmergy specific data model that maps to the components and attributes as illustrated in Figure 1.

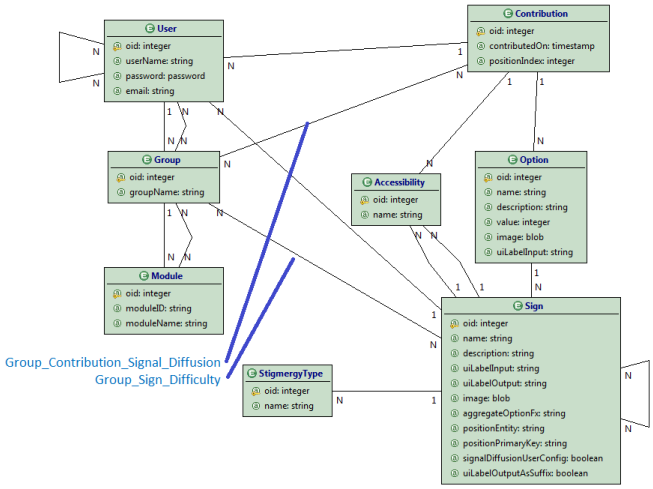


Figure 2. User Accessibility & Stigmergy Positive Feedback Data Model

Our model requires that we add the self-referenced many-to-many relationship to the *user* entity and call the resulting relationship *friends*. It is possible that in future releases this enhancement would become a standard part of WebRatio given current social website networking trends.

The *sign* and *contribution* entities are associated (via Foreign Key) to the *user* and functions as the environmental persistence of the user contributions. The *contribution* is the result of the *option* selected (intentionally) or left (unintentionally) by the user as a byproduct of interaction with the input mechanism. The available options are those which have been defined as being presented for a particular

 ISBN: 978-1-61208-280-6

sign. The *difficulty* entity defines the accessibility for to the *actuator* (or input) to particular users of the site, where *signal diffusion* defines accessibility to the signal *sensor* (or output) through the output mechanisms. The *group* entity is related to the *difficulty* and *signal diffusion* entities (resulting from the many-to-many relationships) defining accessibility and enabling users to set their own privacy levels of contributions. The deployment of this feature is dependent on site-specific implementations. For example, Facebook allows users to restrict the accessibility of their "Share" contributions but not their "Like" contributions. The capability to expose these features for one contribution and not the other are implemented in the *sign* definition as a dynamic feature based the stigmergy data model configuration. Finally, we could consider extending the implementation for *position* to include navigable connected graphs based on Web site hyperlinks but the added complexity is outside the scope of this research project.

The data model supports the definition of signs that fit different varieties of stigmergy. Our quantitative, marker-based example (as illustrated by the Facebook "Like" feature) is defined as a single *sign* record (e.g., of "Like") and a corresponding single *option* record that has a *value* attribute of 1. The sign *aggregationOptionFx* attribute indicates the SUM function for the output mechanism to perform a sum function against to determine the "Like" count. Note: a COUNT function would produce the same result. Given our qualitative, marker-based example we assume that Facebook stores news-feed data in an associated table. The *sign* record defining the sign stores the table name in the *positionEntity* attribute and the name of the primary key field in the *positionPrimaryKey* attribute. This defines the content's position for news feed articles being liked and how that content is referenced. The *signal diffusion* record for the sign is preset to visible to everyone and the *sign.signalDiffusionUserConfigurable* disabled because Facebook does not allow users to set privacy on who can see that they have liked something; anyone with access to the news-feed article can see the "Like" count. The *contribution* record would then store that a particular *user* deposited the "Like" *option* against a news-feed item that is identified by the primary key value stored in the *positionIndex* attribute. The option should be presented as a single button (or hyperlink) being that there is only a single option in this instance, and is labeled using the string stored in the *sign.uiLabelInput* attribute. The output mechanism queries the stigmergy data and presents the result labeled using the string stored in the *sign.uiLabelOutput* attribute. The differentiation of input and output labels is purely for semantics (e.g., Facebook input is "Like" where the output aggregation is "Likes"). The query is a simple sum function of the *contribution.value* where the *positionIndex* is equal to the current news-feed entry's primary key.

If we consider the quantitative, marker-based, eBay example for seller-reputation feedback, there is an input mechanism that allows the selection of a single *option* from multiple options defined against the *sign*. The input mechanism in our proof-of-concept implementation presents a number of options within a drop-down list; however it could also be presented as a radio-button group or a group of buttons / hyperlinks. This *option* presentation is designed into our generic WebRatio output mechanism component and should be dynamic in its ability to render itself according to whether one or many *options* are available for the *sign*. The storage of the contribution remains the same, as does the query used within the output mechanism. A slightly more complex query is required where a composite sign (sign made up of signs) has been defined. The result set is driven by a recursive tree-walk of the *sign* entity generating the collection of *contribution* data grouped by the *positionPrimaryKey* attribute. This applies the sign's *aggregationOptionFx* attribute named function against the collection of children sign's *option.value* attribute where multiple contributions exist. This functionality is hidden in the output mechanism and is transparent to both the user and developer. In the case of eBay where feedback appears to be an average or moving average, the output mechanism can be extended enabling the customization of the aggregation function; however the incorporation of more complex though standard functions can easily be included in the default output mechanism.

The Facebook "Share" feature is an example of qualitative, marker-based stigmergy and follows the same pattern where there is single *option* selection that is associated (via Foreign Key described by meta-data in the *sign* entity) against each news-feed article. Corollary the same data for the *contribution* would be stored; however the difference here is that the user is capable of specifying a different visibility level in *signal diffusion* for their individual *contribution* because that feature is offered to the user. The difference with the output mechanism is that the "Share" feature is provided as part of the core Facebook functionality. The sharing of a news-feed item means that it becomes accessible to the subset of users to whom the content has been shared with. The site functionality provides a qualitative listing (rather than quantitative aggregation function) to exploit this particular signal type. In this example the result set is driven by a query selecting data that is visible to the current user where the contributing *user* is related to them as defined by the *friend* entity relationship or *group* entity which they belong to.

The Facebook "Seen By" feature is an example of sematectonic stigmergy and follows the same pattern where a single *option* is stored as the *contribution*. This is the same *sign* and *option* configuration as outlined in the Facebook "Like" and "Share" examples. The only difference is that the user when navigating a hyperlink to specific content triggers the input mechanism unintentionally. The *option* record for the "Seen By" *sign* has a *value* attribute of 1 and has the same results as for the "Like" *sign* count. The input mechanism is associated to a hyperlink with the pre-defined *option* specified for the

 ISBN: 978-1-61208-280-6

contribution. The output mechanism in this example performs the same sum function of the *contribution.value* and where the *positionIndex* is equal to the current news-feed entry's primary key.

WebRatio provides a modeling interface for design a web application. Predefined components are provided which perform the presentation and transactional operations of a Web site. While stigmergy can be incorporated into a site as a design pattern the optimal approach is to provide reusable components performing the *actuator* and *sensor* dynamics of our model thereby providing reusable input and output mechanisms. Figure 3 provides an example of how an *actuator* can be built using standard WebRatio components based on our design pattern.

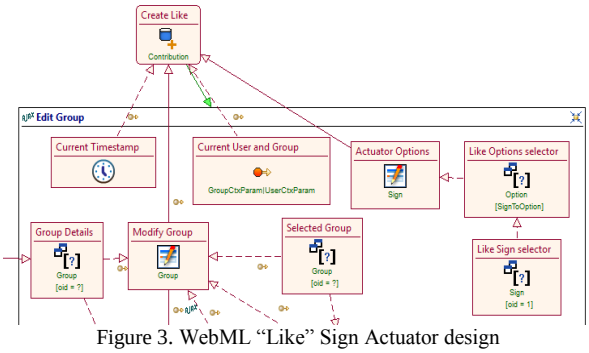


Figure 3. WebML "Like" Sign Actuator design

The example illustrates one way to implement the "Like" sign (predefined within the sign data model entity at primary key oid 1) against a particular group. The *actuator* is located on the page where group entities can be edited and in this example is provided by the link from the Entry Unit named "Modify Group" to the Create Unit named "Create Like". The standard Create Unit is used to insert a record into the contribution entity and the row values area passed in through the links providing values for *contributedOn*, *currentUser*, *currentGroup* and the *option* entity foreign key. The *positionIndex* value is provided using the Link for the currently selected group data entry unit. The specific option to add as the contribution is provided by the Entry Unit named "Actuator Options". The option value is restricted using a Relationship Condition between the "Like Sign selector" and the "Like Options selector". NOTE: In the "Like" stigmergy example only a single option is presented to the user requiring the Entry Unit being configured to not be visible. For examples where multiple options are provided, a Selection Field must be included to present the alternate options within a drop-down list.

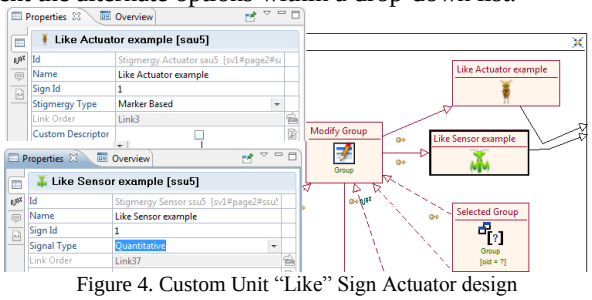


Figure 4. Custom Unit "Like" Sign Actuator design

Figure 4 shows the same example when using our *actuator* Custom Unit within the WebML design. Also shown is our *sensor* Custom Unit. The dramatic simplification is obvious where current user and group values can be obtained within the web service context instead of model links; only the *positionIndex* value needs to be provided via a link. The primary key defining the options for the sign is specified as parameter of the unit, as shown in the Properties window. There is a web service associated with the unit providing database transactions, and more sophisticated algorithms pertaining to the *difficulty* and *signal diffusion* facets of the signal. The final runtime output is displayed in Figure 5.

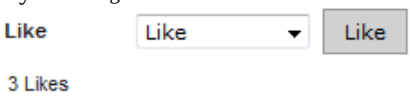


Figure 5. Runtime "Like" *actuator* and *sensor* output

VI. PROBLEMS ENCOUNTERED

WebRatio is a relatively immature product with a small user group and support based. As such there have been a number of problems and impediments encountered.

Custom Units within WebRatio allow definition of page template content to be customized (using scripts) at page generation time, while mark-up tags enable dynamic content during page population at runtime. The limited available tutorials for WebRatio have impeded our progress in optimizing our *actuator* and *sensor* units' implementation. A specific example can be seen in Figure 5 where the runtime *actuator* instance displays a drop-down list containing "Like" and also a button labeled "Like". Our optimal design requires the capacity at page generation time to determine whether a specific sign has a single option, or multiple as defined in the *option* entity. Based on the result of this database query either a single button or a button with the drop-down list would be presented. At present only Custom Unit design-time properties such as the *SignId* as seen in Figure 4 have successfully been prototyped as available at page generation time. A simple work-around for our proof-of-concept has been to split our *actuator* into three separate Custom Units: marker-based (single option), marker-based (multiple options) and sematectonic based. The same applies for our *sensor* design where the Custom Unit split is based on signal type: quantitative and qualitative. Rather than rely on the work-around (which merely means the appropriate Custom Unit be selected for the associated sign data) we pursue the implementation of the consolidated *actuator* and consolidated *sensor* designs. Resolving this issue will also facilitate removing superfluous Custom Unit design-time properties of Stigmergy Type (for *actuator*) and Signal Type (for *sensor*), as they are also defined within the data model. Ideally we will succeed in accessing the stigmergy data model content to define these page qualities at page generation time. However, if we realize that this is not achievable within the WebRatio tool it does not impinge on the correctness of our

 ISBN: 978-1-61208-280-6

model or future experiments on the efficacy of including stigmergy in Web sites.

Since development of the *actuator* and *sensor* prototypes, we have reflected on potential design deficiencies when considering stigmergic mechanisms deployed against our case study Web sites. We have identified two areas for improvements: the ability to revoke a contribution (such as "Unlike" within Facebook); restricting accessibility to the *actuator* after a single contribution (such as providing transaction feedback within eBay). Both of these enhancements can easily be achieved by adding Boolean fields *userRevokable*,(as Boolean) *uiLabelRevoke* (as String) and *singleContrbution* (as Boolean) to the sign entity. The *singleContribution* value will define additional accessibility to the *actuator*, and will provide alternative labeling for revocable contributions.

Progress so far only covers the user-centric dynamics (e.g., *actuate* and *sense* making) of stigmergy that provides the positive feedback system. To fully exploit stigmergy we must also implement the negative feedback illustrated as *atrophy* and *entropy*. WebML and WebRatio provide system events to trigger actions that can drive these environment-centric dynamics. Inclusion of these mechanisms into our proof-of-concept will complete our implementation by introducing the balancing negative feedback of stigmergy. We have yet to address some attributes presented in Figure 1: Progress, Goal, Completion Point, and Significant Dimensions. These attributes pertain to user-centric data (e.g., stored in users heads) and as such are arbitrary as to whether sites facilitate the recording and inclusion of such data. We anticipate addressing these issues in the future.

VII. Conclusion and Future Work

Stigmergy can be seen throughout entomological, human and Web environments. Stigmergy provides a set of dynamics that facilitate a balance between positive and negative feedback within a system. Previously we have presented papers defining *what* stigmergy is. This paper presents *how* the positive feedback mechanisms of stigmergy can be architected into web sites incorporating reusable User Interface components. We have designed a data model supporting each stigmergy variety. Our generic implementation of input and output mechanism within WebRatio demonstrates the simplest examples of stigmergy.

We continue to refine our implementation of positive feedback mechanisms by addressing problems encountered during initial prototype development. Immediate future work is required to consolidate our sensor and actuator units. This requires resolving whether our chosen development environment is technically capable of allow database querying at page generation time, and not solely page population time. We intend to extend model functionality where a Web site design enables user revocation of a contribution or provides a restriction to single, irrevocable contributions. Further model validation will occur by developing prototypes encompassing stigmergy examples observed in alternative case study sites thereby, extending on our Facebook-centric examples. Finally we must extend this data model and include mechanisms that provide negative feedback.

Our proof-of-concept will be used to create an experimental Web site testing user interaction. The analysis will determine how best to employ stigmergy in site designs and assess if stigmergy improves user coordination.

References

[1] G. Theraulaz and E. Bonabeau, "A Brief History of Stigmergy," *Artificial Life,* vol. 5, pp. 97-116, 1999/04/01 1999.

[2] A. Dipple, "Stigmergy in Web 2.0: a Model for Site Dynamics," in *ACM Web Science 2012*, Evanston, 2012, pp. 116-124.

[3] A. Dipple, K. Raymond, and M. Docherty, "Extending Web Modeling Language to Exploit Stigmergy: Intentionally Recording Unintentional Trails," in *WEB 13, The First International Conference on Building and Exploring Web Based Environments*, Seville, Spain, 2013, pp. 87 - 92.

[4] S. Ceri and P. Fraternali, "Model for the definition of world wide web sites and method for their design and verification," US6591271 B1, 2003.

[5] H. Van Dyke Parunak, "A Survey of Environments and Mechanisms for Human-Human Stigmergy," in *Environments for Multi-Agent Systems II*, ed, 2006, pp. 163-186.

[6] Z. Mason, "Programming with stigmergy: using swarms for construction," in *ICAL 2003: Proceedings of the eighth international conference on Artificial life*, 2003, pp. 371-374.

[7] P.-P. Grasse', "La reconstruction dun id et les coordinations interindividuelles chez Bellicositermes natalensis et Cubitermes sp., La theorie de la stigmergie: Essai d'interpretation du comportement des termites constructeurs," *Insectes Sociaux,* vol. 6, pp. 41-80, 1959.

[8] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*. New York: Oxford University Press, Santa Fe Institute Studies in the Sciences of Complexity, 1999.

[9] M. Baumgarten, K. Greer, M. D. Mulvenna, K. Curran, and C. Nugent, "Utilizing Stigmergy in Support of Autonomic Principles," in *Third International Conference on Semantics, Knowledge and Grid*, 2007, pp. 98-103.

[10] E. O. Wilson, *Sociobiology: The New Synthesis*, Twenty-Fifth Anniversary Edition ed.: Belknap Press of Harvard University Press, 2000.

[11] L. Tummolini and C. Castelfranchi, "Trace signals: the meanings of stigmergy," in *E4MAS'06 Proceedings of the 3rd international conference on Environments for multi-agent systems III* Hakodate, Japan, 2007, pp. 141-156.

[12] L. Tummolini, C. Castelfranchi, A. Ricci, M. Viroli, and A. Omicini, "'Exhibitionists' and 'Voyeurs' Do It Better: A Shared Environment for Flexible Coordination with Tacit Messages," in *In 1 st International Workshop on Environments for Multiagent Systems. LNAI, n.3374*, ed, 2004, pp. 215-231.

[13] M. den Besten, L. Gaio, A. Rossi, and J.-M. Dalle, "Using Metadata Signals to Support Stigmergy," in *SASOW '10 Proceedings of the 2010 Fourth IEEE International Conference on Self-Adaptive and Self-Organizing Systems Workshop* 2010, pp. 131-135.

 ISBN: 978-1-61208-280-6

Incorporating Flow Theory to Technology Acceptance Model for Online Community Formation

Mayank Sharma, Pradeep Kumar, Bharat Bhasker and Abhijit Bhatacharya
Indian Institute of Management, Lucknow,
Lucknow,India
{fpm10012,pradeepkumar,bhasker,abhijit}@iiml.ac.in

Abstract - **Online communities in social networking sites exist and thrive in the environment provided by them. The study of online communities becomes important for marketers, academicians and business practitioners because of the huge rise in the user base of these social networking sites in recent years. In this study, we adopt the approach to focus on the user's perceptions about online communities and establish various relationships among them leading to predict the intention to use such online communities. Technology Acceptance Model and flow theory are used to identify the user beliefs affecting the intention to use online communities. We also study the effect of online trust in form of system trust and interpersonal trust on the intention to use online community. We further incorporate the knowledge of user beliefs to establish the benefits in form of online social capital resulting from use of such online communities.**

Keywords-online social network; online communities; TAM; social capital

I. INTRODUCTION

Social networking on the Internet is becoming important not only for individuals but also for organizations. It has penetrated every phase of people's lives from making new friends, to maintaining existing relationships, job search or brand building exercise for organizations. The online social networks (OSNs) provide a shared communication environment with inbuilt features which facilitates the formation of online communities (OC). The study of online communities in social networks thus becomes an integral part of social network researchers.

In today's digital era, information and communication technologies (ICTs) act as a major player or catalyst in the processes of community formation. The communities formed online vary vastly from one another either based on age, culture, economic benefits, interest, language, and other dimensions that would hinder, if not prohibit, communication in the physical world. There could be various reasons for a user to join and remain in a particular online community. These community formations give rise to fundamental questions like, why do people join a community? What factors motivate an online user to share his views and ideas on social networking sites? Do the features like trust and security stops someone from joining a particular online community? How do these variables change over time as the members of a community become more informed about each other?

These issues have been dealt in various studies in different forms like dealing with identity of user in online social networks [1], defining the communities [2], the relative strength of relationship ties among community members, the trust between the community members [3] etc. In this study, we try to address the fundamental aspect of understanding the reasons behind joining an online community.

The objective of our study is twofold in nature. The first objective is to use theories such as Technology Acceptance Model (TAM) [4, 5] and flow theory [6] to identify the factors which can be important determinants for joining an online community. The second objective of our study is to identify the benefits derived by an individual from the online community formation process. We use social capital theory in the form it is applied for online information systems.

In order to accomplish our study, we formulate our research questions as given below:

RQ1: Why do people join online communities? What are the underlying reasons for an individual to join an online community?

RQ2: What are the benefits associated with people joining online communities in context of social networking sites?

The paper is organized as follows. In Section II we present the theoretical background study. In Section III, we present the methodology followed by discussion in Section IV. Finally, we conclude in Section V with the future scope of the study.

II. THEORETICAL BACKGROUND

The objective of this research is to understand the underlying factors which can influence the user to join a particular community, thereby facilitating the community formation process in an online social network. To investigate these factors we start with technology adoption studies and thus use the most widely used Technology Acceptance Model (TAM) [4, 5].

The approach of using theories such as TAM to answer our research question listed in section I, provides us with the user perspective for online community formation process. As suggested by Legris et al. [4] TAM seldom explains the whole picture for technology acceptance problems. So along with TAM we also use factors from flow theory [6], social capital [7] and trust [8] to investigate our research questions more holistically. We build our research model based on the factors from these proven theoretical backgrounds.

A. Intention to use an online community

Technology adoption studies' usual focus is to investigate the intention to use a particular information system and thereby facilitate and translate that intention into

 ISBN: 978-1-61208-280-6

actual use of the information system. This factor becomes the consequence in our research model. It is very important for organizations to know it because it would provide them a way to harness these online communities to reach their organizational goals. Also for marketers it is essential to know the factors impacting the use of an online community.

B. Bonding and Bridging Social Capital

To investigate the benefits associated with joining an online community we use the social capital theory. Social capital broadly refers to the resources accumulated through the relationships among people [10] which provide us two benefits in online social networking sites context. These two benefits are bonding social capital and bridging social capital associated with the type of relationships from which they are derived from online social networks.

C. Perceived ease of use and Peceived usefulness

According to TAM, the attitude towards a technology affects its use. The two belief variables used in TAM which explains the impact on attitude and in turn determine the intention to use, are Perceived Ease of Use (PEOU) and Perceived Usefulness (PU) [23]. As TAM has been applied to online websites or computers at workplace, the following hypotheses follows

Hypothesis 1a: Perceived ease of use will positively affect intention to use an online community.

Hypothesis 1b: Perceived usefulness will positively affect intention to use an online community.

Davis et al. [12] suggest that perceived ease of use affects perceived usefulness and thereby affects the intention to use an information system. Thus follows the hypothesis

Hypothesis 1c: Perceived ease of use will positively affect perceived usefulness of an online community.

As people find merit in perceived ease of use and perceived usefulness of an online community, the more likely they are to use that community and thereby make new ties with other users. The weaker ties contribute to the bridging social capital and stronger ties to the bonding social capital. Thus, we hypothesize

Hypothesis 1d: Perceived ease of use will positively affect bonding social capital.

Hypothesis 1e: Perceived ease of use will positively affect bridging social capital.

Hypothesis 1f: Perceived usefulness will positively affect bonding social capital.

Hypothesis 1g: Perceived usefulness will positively affect bridging social capital.

D. Perceived enjoyment

Past studies have verified that the use of computer technology was influenced by perceived enjoyment (PE) [13]. The concept of perceived enjoyment is borrowed from flow theory [6] where *flow* is defined as "the holistic sensation that people feel when they act with total involvement". A common measure of flow is the level of intrinsic enjoyment of an activity. An online community interaction is a volitional activity where a factor such as enjoyment is likely to play an extremely important role. Thus, we hypothesize

Hypothesis 2a: Perceived enjoyment will positively affect intention to use an online community.

Intrinsic motivation drives a user for joining an online community voluntarily. Therefore, the more a user of a social networking site enjoys the activities within an online community, the more likely he/she will form new ties and thus contribute to social capital. Again the weaker ties contribute to the bridging social capital and stronger ties to the bonding social capital. So the perceived enjoyment resulting from such ties can be hypothesized as follows

Hypothesis 2b: Perceived enjoyment will positively affect bridging social capital

Hypothesis 2c: Perceived enjoyment will positively affect bonding social capital

Venkatesh [14] argues that perceived ease of use is affected by perceived enjoyment. Further Venkatesh [14] found that by manipulating the perceived enjoyment associated with information system not only increased the perceived ease of use of information system but also it became more salient to the intended use of system. Thus, we hypothesize

Hypothesis 2d: Perceived enjoyment will positively affect perceived ease of use of an online community.

E. System trust and Interpersonal trust

Trust is a multidimensional construct whose causes and effects have been studied in various scientific disciplines such as sociology, psychology, and marketing. Trust in online context has been studied by Friedman [8], as a means for enriching social capital, while Ba [3] and Lu et al.[22] studied trust involved with e-commerce transactions etc.

In online community context the trust is derived from the relationships existing among users. This becomes the part of our research model in form of *interpersonal trust*. Interpersonal trust is defined as "an expectancy held by an individual or a group that the word, promise, verbal, or written statement of another individual or group can be relied on" [15]. Another dimension for trust in online community to be considered is the trust of an individual on the overall system. This type of trust is attributed to *system trust* which is defined as perceived integrity, benevolence, and ability of the system operator which in our case is any social networking site.

Benlian and Hess [16] establish trust as an important antecedent for the participation in online communities. The greater trust we have in an online community of a social networking site and its users the more likely we are to join that community and use it. Thus, we hypothesize

Hypothesis 3a: System trust will positively affect intention to use an online community.

Hypothesis 3b: Interpersonal trust will positively affect intention to use an online community.

Since both system trust and interpersonal trust factors into the intended use of an online community, thus they are more likely to form new ties with other users in an online community. Since system trust depends on the individual

 ISBN: 978-1-61208-280-6

perceptions of the institutional environment of a system and the structural assurances it provides, thus it is likely to affect both formation of weak ties and strong ties hence, we hypothesize

Hypothesis 3c: System trust will positively affect bonding social capital

Hypothesis 3d: System trust will positively affect bridging social capital

On the other hand interpersonal trust is a result of interaction among user and is an experience-based trust. Interpersonal trust is contributed to by interaction among users which is more likely to be for the strong ties formed between close friends and family. Thus interpersonal trust would affect the bonding social capital. Thus, we hypothesize as

Hypothesis 3e: Interpersonal trust will positively affect bonding social capital.

F. Social networking site usage, self efficacy and social influence

We now elaborate on the external variables for our research model which are likely to be the antecedents for the user beliefs mentioned earlier.

Eastin and LaRose [17] studied the effect of Internet use on social, informational and entertainment outcomes. More recently Facebook usage has been studied as an antecedent in building of online social capital [9]. We study the usage in our context as social networking site usage (SNS usage) and its effect on the user beliefs to use an online community. Thus in our context of online community interaction we believe that more SNS usage is likely to contribute to greater familiarity with the features of SNS and hence should impact user beliefs to use an online community. Thus, we hypothesize as

Hypothesis 4a: Social networking site usage will positively affect perceived usefulness of an online community.

Hypothesis 4b: Social networking site usage will positively affect perceived ease of use of an online community.

Hypothesis 4c: Social networking site usage will positively affect perceived enjoyment of an online community.

Self-efficacy (SE) in the context of information system adoption studies has been studied extensively [17, 18] as an antecedent [17] to the user beliefs involved in the adoption theories. Self-efficacy is a behavioral concept which was first proposed by Bandura [18] and is defined as the belief "in one's capabilities to organize and execute the courses of action required to produce given attainments". Earlier it has been studied in information system literature as computer self-efficacy and Internet self-efficacy [17]. We, in the context of our study of online communities in social networking sites, use it as social networking site self-efficacy (SNS self-efficacy). Again Internet self-efficacy has been studied as an antecedent for the social, informational and entertainment outcomes [17] and is an important determinant of perceived ease of use of an information system. Liu et al. [19] found that previous online learning experience (an external variable for TAM) is an antecedent for user beliefs such as perceived usefulness and perceived ease of use of TAM. Since SNS self-efficacy relates to the ability and skills that an individual possesses to use a social networking site which can contribute positively to its usefulness, ease of use and enjoyment of an online community. Thus, we hypothesize as

Hypothesis 5a: Social networking site self-efficacy will positively affect perceived usefulness of an online community.

Hypothesis 5b: Social networking site self-efficacy will positively affect perceived ease of use of an online community.

Hypothesis 5c: Social networking site self-efficacy will positively affect perceived enjoyment of an online community.

Other users' actions and thoughts may sometimes influence the choices involved in our decision making. Thus social influence becomes extremely important in social networking studies. Venkatesh et al. [20] define social influence as 'the degree to which an individual perceives that important others believe he or she should use the new system'. Social influence, sometimes termed as subjective norm, has been used as a direct determinant of behavioural intention to use an information system [20]. Venkatesh et al. [20] argues that this relation exists when use of information system is mandated and thus the relation's existence is driven by the compliance factor of social influence. In case of voluntary context, such as use of online communities existing in social networking sites, the social influence exists by virtue of influencing the perceptions about the technology [20]. Thus, we hypothesize

Hypothesis 6a: Social influence will positively affect perceived usefulness of an online community.

Hypothesis 6b: Social influence will positively affect perceived ease of use of an online community.

Hypothesis 6c: Social influence will positively affect perceived enjoyment of an online community.

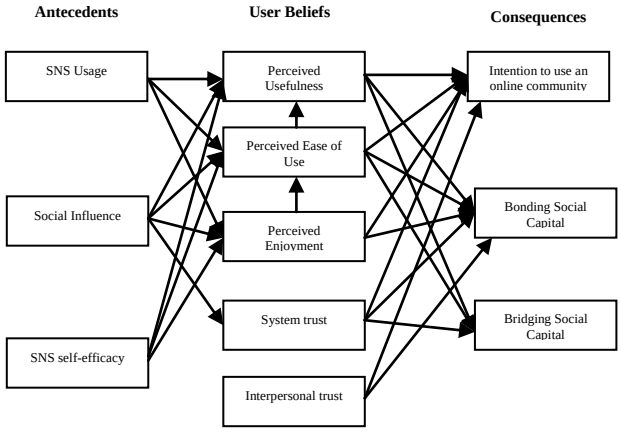


Figure 1. Theoretical research model.

Since social influence in voluntary context can influence the perceptions about an online community, hence it can also contribute in building of trust of users in an online community. Thus, we hypothesize

Hypothesis 6d: Social influence will positively affect system trust of an online community.

The complete theoretical framework is shown in Fig. 1 with the arrows representing the hypothesized relationships between the various constructs.

III. METHODOLOGY

For capturing the user's perception we built a survey instrument based on measurement scales borrowed from the literature. While the measures are based on previously validated instruments in the literature, the current study re-validates these measures.

Both online and offline modes were used for collecting the responses from the survey respondents. A total of 132 responses were collected after rejecting incomplete and invalid responses. For online mode a survey was created on an online survey hosting site and the link was sent with emails to the respondent explaining the nature of study. Each respondent was asked to fill the survey if they had any prior experience of the Internet and also any online community in any social networking site. Participation in the survey was voluntary so that there are no confounding effects from coercing subjects into participation. They were also asked to mention the online community with which they mostly associate with and were part of. Most of the respondents were users of Facebook owing to its largest reach among social networking users.

Table 1 summarizes the reliability and validity of all the scales used in our research model. The average variance extracted (AVE) for every construct was above 0.5, which means the scales had a good convergent validity [21]. We used composite reliabilities (CRs) to evaluate the internal consistency of the measurement model. As shown in Table I, the CRs were all above 0.7, indicating the scales had good reliabilities. All Chronbach's alpha values were above the 0.70 threshold, indicating that the scales had high reliabilities. As Intention to use an online community was measured using single item scale hence the Chronbach's alpha for it cannot be calculated. Also, since social influence is a two item scale we calculate the Spearman-Brown statistic for it, which is 0.742 indicating a good reliability.

We show the correlation matrix and the square roots of the AVEs in Table II. The square roots of the AVEs are the diagonal elements and they were all larger than their corresponding correlation coefficients with other factors. This suggests that the scales had good discriminant validity.

TABLE I. SUMMARY STATISTICS AND CRONBACH'S ALPHA VALUES FOR ALL SCALES

Scale	*Chronbach's alpha*	*AVE*	*CR*
Social networking site usage	0.799	0.910	0.805
Social networking site self-efficacy	0.879	0.935	0.882
Perceived Ease of Use	0.919	0.902	0.906
Perceived Usefulness	0.813	0.796	0.812
Perceived Enjoyment	0.827	0.905	0.908
System Trust	0.897	0.682	0.895
Interpersonal Trust	0.843	0.842	0.821
Bonding Social Capital	0.855	0.862	0.780
Bridging Social Capital	0.919	0.899	0.869

To test the hypotheses we conducted linear regression for each of the dependent variables. The results are shown in Table III and Table IV.

Table III shows the three linear regressions with dependent variable as intention to use an online community, bonding social capital and bridging social capital respectively.

The first linear regression in Table III has intention to use an online community as the dependent variable and user beliefs such as perceived usefulness, perceived ease of use and perceived enjoyment as independent variables. Along with them, system trust and interpersonal trust are also

TABLE II. CORRELATION MATRIX AND SQUARE ROOTS OF AVE

Scale	SNS Usage	SE	PE	IT	BOSC	BRSC	ST	PU	PEOU	IU	SI
SNS Usage	**0.954**										
SE	0.267**	**0.967**									
PE	0.453**	0.494**	**0.951**								
IT	0.209*	0.298**	0.229**	**0.918**							
BOSC	0.278**	0.451**	0.386**	0.416**	**0.928**						
BRSC	0.474**	0.549**	0.491**	0.157	0.349**	**0.948**					
ST	0.335**	0.455**	0.476**	0.625**	0.552**	0.377**	**0.826**				
PU	0.386**	0.554**	0.498**	0.221*	0.455**	0.598**	0.410**	**0.892**			
PEOU	0.200*	0.497**	0.510**	0.254**	0.386**	0.356**	0.386**	0.565**	**0.950**		
IU	0.214*	0.398**	0.264**	0.113	0.225**	0.370**	0.285**	0.539**	0.337**	N/A	
SI	0.217*	0.343**	0.374**	0.302**	0.456**	0.392**	0.602**	0.506**	0.409**	0.366**	N/A
Mean	3.32	3.56	3.65	2.89	3.08	3.65	3.29	3.55	3.86	3.53	3.32
Variance	0.549	0.419	0.387	0.509	0.411	0.421	0.435	0.395	0.333	0.640	0.485

Notations used above: SNSU - Social networking site usage; SE- Self-efficacy; PE= Perceived enjoyment; IT- Interpersonal trust; BOSC- Bonding social capital; BRSC- Bridging social capital; ST- System trust; PU-Perceived usefulness; PEOU- Perceived ease of use; IU- Intention to use ; SI- Social Influence

 ISBN: 978-1-61208-280-6

independent variables. This model has a good fit and explains variance with $R^2 = 0.303$. With this result we infer that perceived usefulness was the most important determinant for intention to use an online community. Perceived ease of use and perceived enjoyment are not found to be significant because they are indirectly influencing intention to use an online community through perceived usefulness. System trust and Interpersonal trust are also not found to be statistically significant. Hence only hypothesis H1b is supported but hypotheses H1a, H2a, H2b, H3a and H3b are not supported.

TABLE III. LINEAR REGRESSION FOR INTENTION TO USE AN ONLINE COMMUNITY, BONDING SOCIAL CAPITAL AND BRIDGING SOCIAL CAPITAL

Linear Regression for	**Variable**	***B***	***Std. error***	***β***
Intention to use an online community $R^2 = 0.303; F = 10.939$ ($p < 0.01$)	Constant	0.981	0.462	
	Perceived Usefulness	0.637	0.122	0.500**
	Perceived Ease of use	0.072	0.133	0.052
	Perceived Enjoyment	-0.077	0.123	-0.60
	System Trust	0.173	0.131	0.143
	Interpersonal Trust	-0.097	0.108	-0.087
Bonding social capital $R^2 = 0.395$; $F = 13.608$ ($p < 0.01$)	Constant	0.316	0.349	
	Perceived Usefulness	0.190	0.095	0.186*
	Perceived Ease of use	0.057	0.101	0.051
	Perceived Enjoyment	0.028	0.093	0.028
	System Trust	0.295	0.099	0.304**
	Interpersonal Trust	0.113	0.081	0.126
Bridging social capital $R^2 = 0.419$; $F = 18.170$ ($p < 0.01$)	Constant	1.032	0.342	
	Perceived Usefulness	0.493	0.090	0.477**
	Perceived Ease of use	-0.081	0.099	-0.072
	Perceived Enjoyment	0.250	0.091	0.240**
	System Trust	0.138	0.097	0.141

where N = 132 ; ** $p < 0.01$, * $p < 0.05$

The second linear regression in Table III shows the linear regression with bonding social capital as the dependent variable and user beliefs such as perceived usefulness, perceived ease of use and perceived enjoyment as independent variable along with system trust and interpersonal trust. This linear regression model has good fit and explains about 39.5% variance. The coefficients for perceived usefulness and system trust are found to be significant but perceived ease-of-use, perceived enjoyment and interpersonal trust are not significant. Hence hypotheses H1f and H3c are supported but hypotheses H1d and H2c are not supported.

The third linear regression in Table III shows the linear regression with bridging social capital as the dependent variable and user beliefs such as perceived usefulness, perceived ease of use and perceived enjoyment as independent variable along with system trust and interpersonal trust. This linear regression model has good fit and explains about 42% variance. The coefficients for perceived usefulness and perceived enjoyment are found to be significant but perceived ease-of-use and system trust are not significant. Hence hypotheses H1g and H2b are supported but hypotheses H1e and H3d are not supported.

Table IV shows the four linear regressions with dependent variable as perceived usefulness, perceived ease-of-use, perceived enjoyment and system trust respectively.

The first linear regression in Table IV shows the linear regression with perceived usefulness as the dependent variable and external variables such as SNS usage, social influence and SNS self-efficacy along with perceived ease-of-use as independent variable. This linear regression model has good fit and explains about 52% variance. The coefficients for SNS usage, social influence, SNS self-efficacy and perceived ease-of-use are found to be significant. Hence all hypotheses related with this linear regression model viz. H1c, H4a, H5a and H6a are supported.

The second linear regression in Table IV shows the linear regression with perceived ease-of-use as the dependent variable and external variables such as SNS usage, social influence and SNS self-efficacy along with perceived enjoyment as independent variable. This linear regression model has good fit and explains variance with $R^2 = 0.377$. The coefficients for social influence, SNS self-efficacy and perceived enjoyment are found to be significant but SNS usage are not significant. Hence hypotheses H2d, H5b and H6b are supported while hypothesis H4b is not supported.

The third linear regression in Table IV shows the linear regression with perceived enjoyment as the dependent variable and external variables such as SNS usage, social influence and SNS self-efficacy as independent variable. This linear regression model has good fit and explains variance with $R^2 = 0.385$. The coefficients for social influence, SNS self-efficacy and SNS usage are found to be significant. Hence all hypotheses related with this linear regression model viz. H4c, H5c and H6c are supported.

The fourth linear regression in Table IV shows the linear regression for system trust as the dependent variable and social influence as independent variable. This linear regression model has good fit and explains variance with $R^2 = 0.362$. The coefficient for social influence is found to be significant hence we conclude hypothesis H6d is supported. Fig. 2 shows the relationships supported in theoretical model by our empirical study.

 ISBN: 978-1-61208-280-6

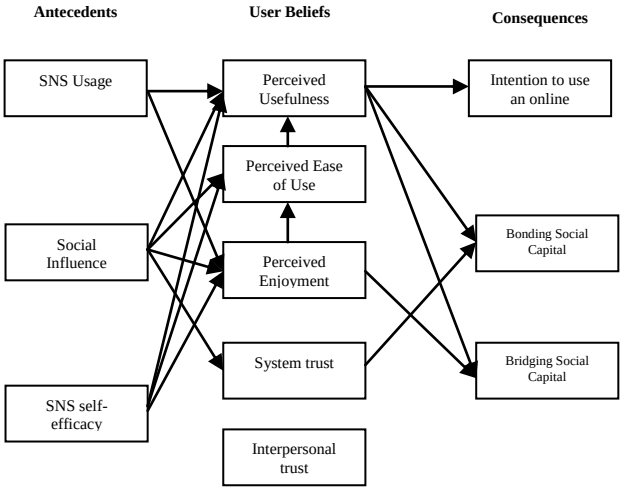


Figure 2. Relationships supported empirically

IV. DISCUSSION

In our study, first we tried to establish the relationships between the external variables, user beliefs and their consequences. The benefits in form of bonding social capital and bridging social capital were also incorporated and the effect of user beliefs and external variables on them was also tested. The results show that intention to use an online community is directly driven by its perceived usefulness, because the more a user finds an online community useful, more likely he/she is to join and use that community. The perceived enjoyment and perceived ease-of-use do not have a direct impact on use of online community but they influence the user's decision indirectly by increasing the usefulness aspect of an online community. This is in line with the voluntary context of online community usage, whereby the perceived enjoyment and ease-of use of online community increases its perceived usefulness. For example, the easier a user finds an online community features, the more useful that community will become for him/her, because he/she would now be able to derive more from the community experience using those features. Therefore, while it may not be expected that an online community may be used for the purpose of enjoyment, but if a user enjoys the experience then it may contribute towards the usefulness of that online community. Also with our study we once again established the robustness of TAM in being able to predict the intention to use an information system. Furthermore, the usage, social influence and self-efficacy of social networking sites influence the user beliefs. Thus, the more time a user spends on a social networking site, the more likely he/she is to gain the ability and skills, and therefore become comfortable with the social networking site features. Also this would help in the formation of

TABLE IV. LINEAR REGRESSION FOR PERCEIVED USEFULNESS, PERCEIVED EASE OF USE , PERCEIVED ENJOYMENT AND SYSTEM TRUST

Linear Regression for	**Variable**	***B***	***Std. error***	***β***
Perceived Usefulness $R^2 = 0.517$; F = 34.004 ($p < 0.01$)	Constant	0.088	0.305	
	SNS usage	0.171	0.055	0.202**
	Social Influence	0.227	0.071	0.251**
	SNS self-efficacy	0.263	0.062	0.271**
	Perceived Ease of use	0.313	0.081	0.287**
Perceived Ease of use $R^2 = 0.377$; F = 19.196 ($p < 0.01$)	Constant	1.471	0.293	
	SNS usage	-0.52	0.061	-0.067
	SNS self-efficacy	0.254	0.073	0.285**
	Social Influence	0.169	0.064	0.204**
	Perceived Enjoyment	0.3	0.083	0.323**
Perceived Enjoyment $R^2 = 0.385$; F = 26.760 ($p < 0.01$)	Constant	1.022	0.299	
	SNS usage	0.269	0.061	0.320**
	SNS self-efficacy	0.332	0.073	0.345**
	Social Influence	0.166	0.067	0.186**
System Trust $R^2 = 0.362$; F = 73.755 ($p < 0.01$)	Constant	1.393	0.225	
	Social Influence	0.570	0.066	0.602**

where N = 132 ; ** $p < 0.01$, * $p < 0.05$

positive beliefs about using an online community. The social influence in voluntary context is driven by internalization and identification component [20] which affects the user perceptions about technology and thus it is seen to be influencing the TAM variables as well as system trust. However we did not find any influence of system trust and interpersonal trust on intention to use an online community. This is in contrast with the expected effect and thus needs to be further tested and investigated. It may be the case that since respondents in this study were majorly college students, they may have known each other through earlier interactions thus they were able to trust the online community and its members. We may test this effect on respondents who are likely to join an online community which is not as much influenced by their offline interactions. Still we found that in terms of benefits of joining an online community the bonding social capital is influenced both by user beliefs about online community as well as the system trust component on the online community. Thus members who are joining an online community and remain with that community for a long time are able to strengthen the system trust component and hence would be able to benefit in terms of their bonding social capital. The strong tie formation with community members would provide emotional support to the user. The

 ISBN: 978-1-61208-280-6

information exchange benefits from a diverse set of users of an online community in form of bridging social capital is seen to be derived directly/indirectly from the user beliefs i.e. TAM variables perceived usefulness, perceived ease-of-use and perceived enjoyment. A user finding the online community useful and enjoyable is likely to develop new relationships with online community members and thus would be exposed to diverse information exchanges contributing to the bridging social capital. However, again, we could not establish the relationship of trust on bridging social capital which thus needs to be investigated further with a more diverse sample of users for our study.

V. CONCLUSION AND FUTURE SCOPE

In this study, we have developed a theoretical model which incorporates the theories such as TAM and flow theory to explain the user behavior in the community formation process. We conclude that usefulness of online community is an important driver for joining an online community. The factors, identified from our research study, such as self-efficacy and social influence can be used towards making an online community useful. We also found that online social capital for both strong and weak ties is built mainly due to the usefulness of an online community. Trust in online communities also strengthen the relationships among users of an online community and thus contributes towards bonding social capital.

Future direction can be to compare the results from our study with results from different type of online communities. This approach can either strengthen or provide more insight into the results from our study.

REFERENCES

[1] A. Marwick, "'I'm a Lot More Interesting than a Friendster Profile': Identity Presentation, Authenticity and Power in Social Networking Services," Association of Internet Researchers, 2005, pp. 6.

[2] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks," Proceedings of the National Academy of Sciences of the United States of America, 101(9), 2004, pp. 2658-2663.

[3] S. Ba, "Establishing online trust through a community responsibility system," Decision Support Systems, 31(3), 2001, pp.323-336.

[4] P. Legris, J. Ingham, and P. Collerette, "Why do people use information technology? A critical review of the technology acceptance model," Information and management, 40(3), 2003, pp.191-204.

[5] F.D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," MIS Quarterly 13(3) 1989, pp.319–340.

[6] M. Csikszentmihalyi, "Play and intrinsic rewards," Journal of Humanistic Psychology, 1975, pp. 41-63.

[7] C.Steinfield, N.B. Ellison, and C. Lampe, "Social capital, self-esteem, and use of online social network sites: A longitudinal analysis," Journal of Applied Developmental Psychology, 29(6), 2008, pp. 434-445.

[8] B. Friedman, P.H. Khan Jr, and C.D. Howe, "Trust online," Communications of the ACM, 43(12), 2000, pp. 34-40.

[9] 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, 12(4), 2007, pp.1143-1168.

[10] J.S. Coleman, "Social capital in the creation of human capital. American Journal of Sociology," 94(Supplement), 1988, pp. S95–S120.

[11] R.D. Putnam, "The prosperous community," The American prospect, 4(13), 1993, pp. 35-42.

[12] F. D. Davis, R.P. Bagozzi, and P.R. Warshaw, "User acceptance of computer technology: a comparison of two theoretical models," Management science, 35(8), 1989, pp. 982-1003.

[13] A. Dickinger, M. Arami, and D. Meyer, "The role of perceived enjoyment and social norm in the adoption of technology with network externalities," European Journal of Information Systems, 17(1), 2008, pp. 4-11.

[14] V. Venkatesh, "Creation of favorable user perceptions: exploring the role of intrinsic motivation," MIS quarterly, 1999, pp. 239-260.

[15] J.B. Rotter, "Generalized expectancies for interpersonal trust" American Psychologist; American Psychologist, 26(5), 1971, pp. 443- 443.

[16] A. Benlian and T. Hess, "The Signaling Role of IT Features in Influencing Trust and Participation in Online Communities. International Journal of Electronic Commerce, 15(4), 2011, pp. 7-56.

[17] M.S. Eastin and R. LaRose, "Internet self-efficacy and the psychology of the digital divide," Journal of Computer-Mediated Communication, 6(1), 2000 pp. 0-0.

[18] A. Bandura, "Self-efficacy: toward a unifying theory of behavioral change," Psychological review, 84(2), 1977,pp. 191-191.

[19] I.F. Liu, M.C. Chen, Y.S. Sun, D. Wible, and C.H. Kuo, "Extending the TAM model to explore the factors that affect Intention to Use an Online Learning Community," Computers and Education, 54(2), 2010, pp. 600-610.

[20] V. Venkatesh, M.G. Morris, G.B. Davis, and F.D. Davis, "User acceptance of information technology: Toward a unified view," MIS quarterly, 2003, pp. 425-478.

[21] R.P. Bagozzi and Y. Yi, "On the evaluation of structural equation models," Journal of the academy of marketing science, 16(1), 1988, pp.74-94.

[22] Y.Lu, L.Zhao, and B.Wang, "From virtual community members to C2C e-commerce buyers: Trust in virtual communities and its effect on consumers' purchase intention," Electronic Commerce Research and Applications, 9(4), 2010, pp. 346-360.

[23] C.Lorenzo-Romero, E.Constantinides, and M. D. C. Alarcon-del-Amo, "Consumer adoption of social networking sites: implications for theory and practice," Journal of Research in Interactive Marketing, 5(2/3), 2011, pp. 170-188.

 ISBN: 978-1-61208-280-6

Creativity Detection in Texts

Costin – Gabriel Chiru
Department of Computer Science and Engineering
Politehnica University of Bucharest
Bucharest, Romania
E-mail: costin.chiru@cs.pub.ro

***Abstract*— In this paper, we present a model that was intended to discriminate creative from non-creative news articles. In order to build the classifier, we have combined nine different measures using a stepwise logistic regression model. The obtained model was tested in two experiments: the first one tried to discriminate between news articles about the US 2012 Elections from different newspapers versus articles taken from The Onion (a website providing satiric news) on the same subject, while the second one evaluated the capacity of the model to generalize over different topics and text genres. The experiments showed that the system achieves 80% accuracy, but the lack of true positives from the second experiment raised the question of whether we really identified creativity or in fact we detected satire (as the assumption for the training corpus was that the satiric news from The Onion were also creative).**

Keywords-Creativity; Satire; Natural Language Processing; Metrics for Creativity Detection

I. INTRODUCTION

According to Zhu et al. [1], the definition of creativity is the ability to transcend traditional ideas, patterns, relationships into meaningful new ideas, interpretations, etc.

The goal of this paper is to identify whether a text is creative or not. To determine this, several steps were undertaken. First of all, we tried to identify the elements that define a creative text. After that, the most important features that explain creativity were chosen. A model for automatic creativity detection was derived as the final result.

In order to do that, nine different measures were explored: Type-to-Token Ratio [2], Word Norms Fraction [2], Google Similarity Distance [3], Explicit Semantic Analysis [4], Number of Named Entities, Named Entities Score, Wordnet Similarity [5], Coherence measure [6], and Latent Semantic Analysis (LSA) measures [7].

The paper continues with a short presentation of the current approaches for creativity detection and after that we describe the nine measures that we have investigated in our experiments for identifying creativity. Section 3 details the architecture of our application and after that we present the two experiments that we undertook and the obtained results. The paper ends with our conclusions based on these experiments.

II. STATE OF THE ART

Renouf [8] describes creativity as the thought of acting or the quality of an unpredictable departure from the rules of regular word formation.

In texts, the creativity measures "new and creative ways of expressing a given idea" and it is called linguistic creativity [9]. Measuring it has been known for its complexity.

A machine learning algorithm has been developed by Zhu et al. [1] to measure creativity by developing subjective creativity metrics. The aim of the algorithm was to use a linear regression model with 17 features derived from computer science and psychology perspectives [1].

Jordanous [10] proposed a Standardized Procedure for Evaluating Creative Systems (SPECS), which follows three steps for determining whether a computational system can be defined as creative or not. The three steps are: creativity identification, the derivation of standards to be used for evaluation of creativity and the system testing according to those standards [10].

Other researches related to linguistic creativity were focused on understanding and using metaphors [11][12] and analogies [12] or on explaining the appearance of new words from already existing ones (e.g., "television"+ "marathon" = "telethon")[13].

Creativity detection in song lyrics has also been carried out by Hu and Yu [2] by comparing three measures, two of them being adapted from the work of Zhu et al. [1]. Those two measures were Word Norms Fraction and Wordnet Similarity. The metrics proved to be able to determine the different aspects of identifying mood and creativity in a lyric. The Word Norms Fraction was used to calculate the lyrics' "usualness", while WordNet Similarity was involved in determining the similarities between concepts [2].

Still, the research for identifying creativity is in its early stages and in this paper we intended to make a step forward by developing a model that is able to discriminate between creative and non-creative texts, using a stepwise logistic regression model built on a corpus of creative and non-creative texts.

III. CREATIVITY MEASURES

In order to decide where creativity occurs, there is a widespread support that two important criteria are *novelty* and *quality*:

- *Novelty*: To what extent an item is different to the existing samples of its genre?
- *Quality*: How good the item really is?

In this paper, we tried to capture these two criteria through nine different measures: some of them were intended to capture novelty by identifying how ordinary a text is, while the others – the semantic ones – were used for detecting the quality of that text. Combining them, we hoped

 ISBN: 978-1-61208-280-6

that we would be able to determine the degree of creativity of a given text. The nine measures that we investigated are described in further detail bellow.

A. Type-To-Token Ratio

Type-to-Token Ratio is defined as the number of unique terms in a text divided by the total number of terms. It is often used to measure the vocabulary richness of a text [2].

$$m_1 = \frac{C_{unique}(x)}{n} \quad (1)$$

where C_{unique} is the number of unique words in a text and n is the total number of words.

B. Word Norms Fraction

Word norms represent associations between words, while Word Norms Fraction measures the "usualness" of a text [2]. According to Hu and Yu [2], texts with high occurrences of word norms should indicate high "usualness" and thus low creativity since creativity often corresponds to unusual patterns. In order to compute the "usualness" of word pairs, we have used the 72,176 pairs of word pairs offered by Free Association Norms [14].

$$m_2 = \frac{C_{norm}(x,y)}{n} \quad (2)$$

where $C_{norm}(x,y)$ is the number of word pairs that appear in Free Association Norms and n is the total number of words in the text.

C. Google Similarity Distance

Google Similarity Distance [3] measures similarity of words and phrases from the World Wide Web using Google page counts. It is based on the concept that the probabilities of Google search terms (conceived as the frequencies of page counts returned by Google divided by the number of pages indexed by Google), approximate the relative frequencies of those search terms as actually used in society. The Google Similarity Distance is given by:

$$m_3 = \frac{\max\{log(f(x)),log(f(y))\} - \log(f(x,y))}{\log(M) - \min\{log(f(x)),log(f(y))\}} \quad (3)$$

where f(x) denotes the number of pages containing x, and f(x, y) denotes the number of pages containing both x and y. M is the total number of web pages indexed by Google and during their experiments (in 2007), it was shown to have a value of 8,058,044,651. Nowadays, M is considered to have a value of 50 billion [15]. This value was used for the calculation of the Google Similarity Distance.

D. Explicit Semantic Analysis

The aim of the Explicit Semantic Analysis (ESA) [4] is to compute the semantic relatedness between the vectors of words using Wikipedia as the knowledge base. Wikipedia has been known as the largest online knowledge repository and it has been proven to be highly organized and regularly maintained, thus ensuring its consistency. This method uses Wikipedia's concepts and explicitly represents the meaning of a given text in terms of the concepts in Wikipedia. ESA manipulates concepts based on human cognition, which is why it is explicit in a sense compared to the Latent Semantic Analysis approach.

The input of this method is a plain text with concepts represented by the Wikipedia articles ranked according to their relevance using classic text classification algorithms. Each concept is represented as an attribute vector with assigned weights using Term Frequency–Inverse Document Frequency (TFIDF) and afterwards an inverted index is built. Once the text is represented by a semantic interpretation vector, simple cosine similarity is used to compute the semantic relatedness.

E. Number of Named Entities

This measure gives the total number of named entities found in the text, in order to check if the creativity of a text is related to the number of named entities used in the text.

m_5 = number of named entities in a text

F. Named Entities Score

This measure gives the proportion of distinct named entities used in the text. It is computed by dividing the number of distinct named entities by the total number of named entities.

$$m_6 = \frac{Number\ of\ distinct\ named\ entities}{Total\ number\ of\ named\ entities} \quad (4)$$

G. WordNet Similarity

The WordNet Similarity measure is based on the lexical database Wordnet [5]. It returns a value denoting how similar two word senses are, based on the shortest path that connects their senses in the WordNet lexical ontology.

H. Coherence Measure

Coherence can be thought of as how meanings and sequences of ideas relate to each other in a text. One approach of measuring coherence in a text is to compare sentences and check how similar they are. The coherence measure we propose is computed from the pair-wise sentence similarity. This measure is based on the coherence score proposed by He et al [6].

Given a set of documents D = $\{d_i\}$, i = 1...M, we define the coherence score as the proportion of "coherent" pairs of documents with respect to the total number of document pairs within D. The criterion of being a "coherent" pair is that the similarity between the two documents in the pair should meet or exceed a given threshold. Formally, given the document set D and a threshold τ, we have:

$$\delta(d_i, d_j) = \begin{cases} 1, & if\ sim(d_i, d_j) \geq \tau \\ 0, & otherwise \end{cases} \quad with\ i \neq j \in \{1..M\} \quad (5)$$

where cosine similarity is taken as the similarity between documents d_i and d_j and the threshold is set to 0.05. Then, coherence score of the document set D is defined as:

 ISBN: 978-1-61208-280-6

$$m_8 = \frac{\sum_{i \neq j \in \{1..M\}} \delta(d_i, d_j)}{\frac{1}{2}M(M-1)} \quad (6)$$

I. LSA Measures

Latent Semantic Analysis (LSA) is the best known and most widely used vector-space method for computing semantic similarity using dimensionality reduction [7]. It involves the application of Singular Value Decomposition (SVD) to a document-by-term matrix to reduce its rank. We used LSA to analyze sentence-to-sentence similarity of texts. Each sentence is treated as a document and LSA is performed on the document-by-term matrix. From the resulting matrix with reduced dimension, four different measures are computed.

1) Average similarity between adjacent sentences

From the reduced dimensionality matrix $\hat{X}$, the sentence similarity matrix S = [s_{ij}] = $\hat{X}^T\hat{X}$ is computed. The matrix S gives the similarity of all pairs of sentences. Average similarity between adjacent sentences is computed as follows:

$$m_{9a} = \frac{\sum_{i=1,j=i+1}^{n-1} s_{ij}}{n-1} \quad (7)$$

2) Average similarity between sentences

From the sentence matrix S, average similarity between pairs of sentences is given by:

$$m_{9b} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}}{\frac{1}{2}n(n-1)} \quad (8)$$

3) Average cosine similarity between adjacent sentences

This measure is similar to m_{9a} and gives the average of similarity of all pairs of adjacent sentences. However, this measure uses cosine similarity instead of the sentence similarity matrix S. Cosine between the sentence-vectors obtained from SVD is computed for all pairs of adjacent sentences and their average is taken.

$$m_{9c} = \frac{\sum_{i=1,}^{n-1} Cosine(\hat{x}_i, \hat{x}_{i+1})}{n-1} \quad (9)$$

4) Average cosine similarity between sentences

This measure is similar to m_{9b} but like in m_{9c}, cosine similarity is used to compute this measure. This measure is the average of cosine between the sentence-vectors obtained from SVD computed for all pairs of sentences.

$$m_{9d} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} Cosine(\hat{x}_i, \hat{x}_j)}{\frac{1}{2}n(n-1)} \quad (10)$$

IV. Experiment's Architecture

The main experiment architecture consisted of three main modules: web crawling, corpus building and creativity assessment using the creativity measures presented above. The first two modules are shown in Figure 1, while the third one is detailed in Figure 2. Each process will be further detailed.

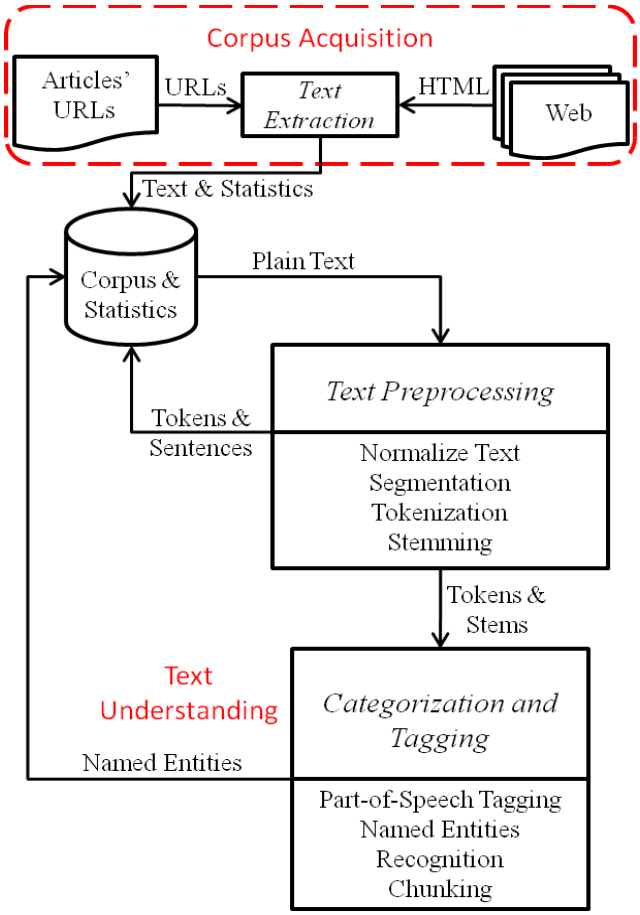


Figure 1. First two modules of the experiment: Web Crawling and Corpus Building involving Text Extraction, Preprocessing, Categorization and Tagging.

A. News articles extraction

During the experiment, we selected 118 articles that were debating about the US 2012 Elections. The articles were collected from 12 news sites from six different countries:

- *UK*: BBC, Wired, The Independent, The Sun;
- *Canada*: CBC;
- *Australia*: News.com.au, The Australian, Sydney Morning Herald;
- *USA*: Foxnews and Huffington Post;
- *South Africa*: News24;
- *New Zealand*: The NZ Herald.

In addition, 67 articles on the same topic were also extracted from The Onion [16]. As The Onion is a satire news organization, these articles were assumed to be more creative than the news. This assumption is based on the fact that while the news articles only present the facts/events, the ones from The Onion should involve either additional feelings towards these facts/events that would transform the articles into satires or satiric parallelisms with other facts/events (otherwise not being published). And both these actions could be triggers for creativity.

 ISBN: 978-1-61208-280-6

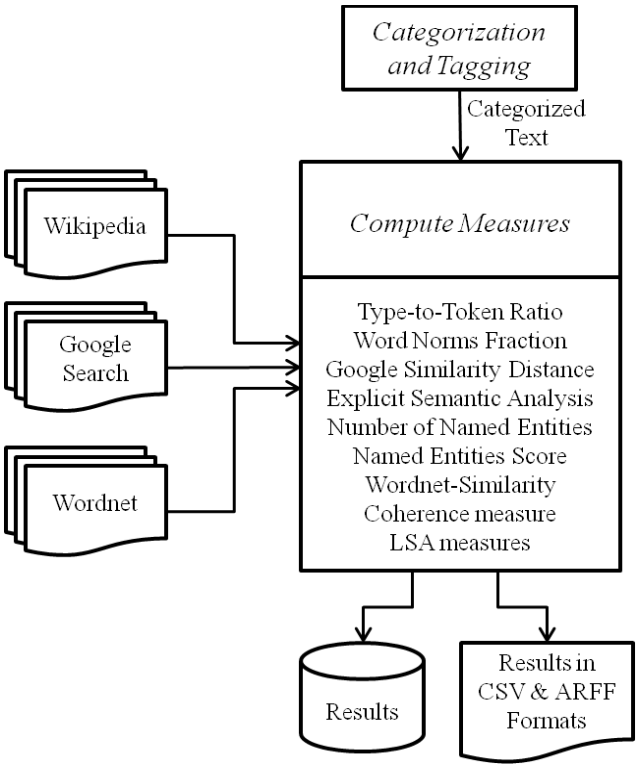


Figure 2. Measures Calculation modules.

B. Preprocessing, Categorization and Tagging

Once the articles' content was saved, preprocessing techniques (from the field of Natural Language Processing) were applied to the text (such as: tokenizing, sentence segmentation, stemming, and stop words removal). The next step was the part-of-speech tagging of the text and after that the assignment of categories (creative or not) based on being a The Onion article or not.

C. Computing the measures for each text

The next process consisted of computing the nine measures described above for each of the gathered documents. The process used information from Wikipedia, Google search and Wordnet (see Figure 2). The obtained results were min-max normalized to set the values in a range between 0 and 1 and then they were saved in order to be used as input files for the predictive models that are built in the next step.

D. Stepwise Logistic Regression

To evaluate the performance of each measure and to obtain a model able to predict the creativity or non creativity of an input text, a stepwise logistic regression model was built. This model gives a weight for each of the used features (which can be interpreted as an importance coefficient), helping to identify those that are most relevant. These weights show how strong is the correlation of the corresponding metric with creativity. The larger the absolute value of the coefficient is, the more important is the feature in determining whether the text is creative or not. This part was done with the help of Orange [17] and Tanagra [18], both open source data mining tools. The diagram for the workflow designed in Orange is shown in Figure 3.

V. Experiments and Results

In order to evaluate our work, we did two experiments: the first one was done for assessing the value of our built classifier and was tested versus the news articles that we have extracted, while the second one was intended to measure the capacity of the classifier to adapt (to what degree the classifier can be generalized in order to be applied to any kind of text?).

A. Assessment Experiment

The first thing that we had to do was to determine the parameters of the logistic regression model based on the values that we obtained for the set of news articles that we extracted from the web. A graphical representation of the values obtained for these documents for each of the measures described in Section 3 is provided in Figure 4.

After applying the logistic regression to this data set augmented with the obtained measures for each of the news articles, the equation defining the creativity (the negative class) or non-creativity (the positive class) was given by the formula:

$$\Pr(Y = 1 | X1, \dots, X9) = F(Bi * Xi),\ i = 0..9 \qquad (11)$$

, where: $X_0 = 1$, $X_1 - X_9$ represent the measures (Type-To-Token Ratio, Word Norms Fraction, Google Similarity Distance, ESA, Number of Named Entities, Named Entities Score, Wordnet Similarity, Coherence Measure, LSA), while B_0 is the bias factor and $B_1 - B_9$ are the parameters associated to each of the measures. The values obtained from the model were: $B_0 = 1.83$, $B_1 = 0$, $B_2 = 3.585$, $B_3 = 3.255$, $B_4 = 0$, $B_5 = -2.897$, $B_6 = 2.485$, $B_7 = -9.799$, $B_8 = 0$, $B_9 = 3.445$, resulting in a classifier for being creative or not, given by (12).

$$\Pr(Y = 1 \mid X_1, \dots, X_9) = F(1.83 + 3.585 * X_2 + 3.255 * X_3 - 2.897 * X_5 + 2.485 * X_6 - 9.779 * X_7 + 3.445 * X_9) \qquad (12)$$

A couple of observations should be drawn based on the model represented by (12). First of all, one can see that the bias factor (B_0) is positive, reflecting the fact that most of the texts are non-creative. Secondly, we saw that the Type-To-Token Ratio had no influence against the creativity of a text. This implies that both creative and non-creative texts had similar ratios of unique terms. On the other hand, Word Norms Fraction had the highest positive influence (showing evidence of a non-creative text), which was confirming our expectations since high values for this measure witnessed high "usualness" of the text, which contradicts the definition of creativity. More than that, the high value received by the parameter of the Google Similarity Distance comes to augment the drive towards the text "usualness".

Regarding the analysis of named entities, the classifier considered that the use of named entities is a sign of creativity (the parameter for the Number of Named Entities

 ISBN: 978-1-61208-280-6

is negative), but in the same time it regards the use of distinct such entities as being non-creative (positive Named Entities Score) which was at least confusing at the beginning. After a deeper analysis of the texts, we have reached the conclusion that this fact was correct, since the more distinct named entities were found in text, the less space was dedicated to expressing the author's sentiments related to the events described (which we consider to offer the opportunity for creativity) because that space was filled with the facts expressed by the named entities.

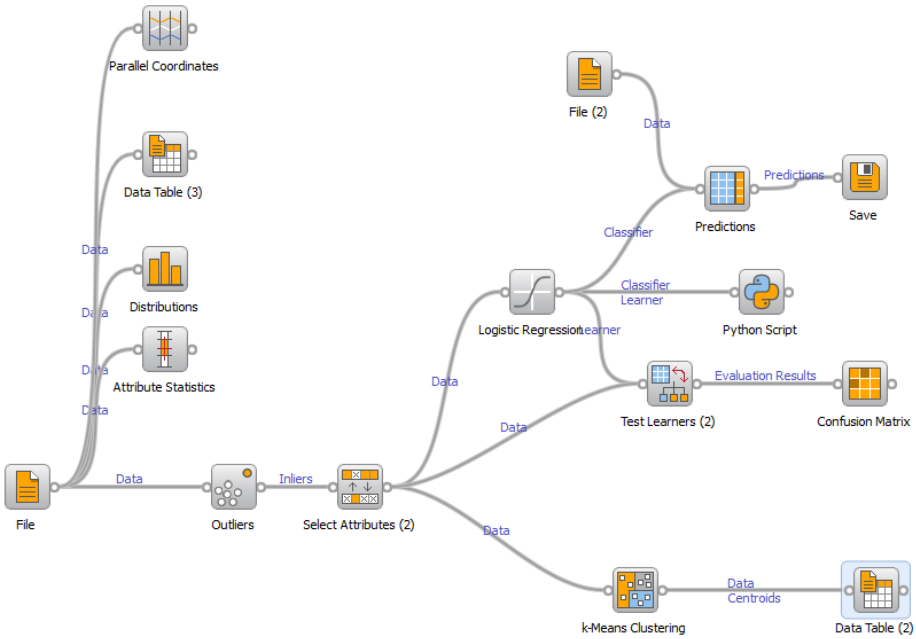


Figure 3. Orange data analysis workflow

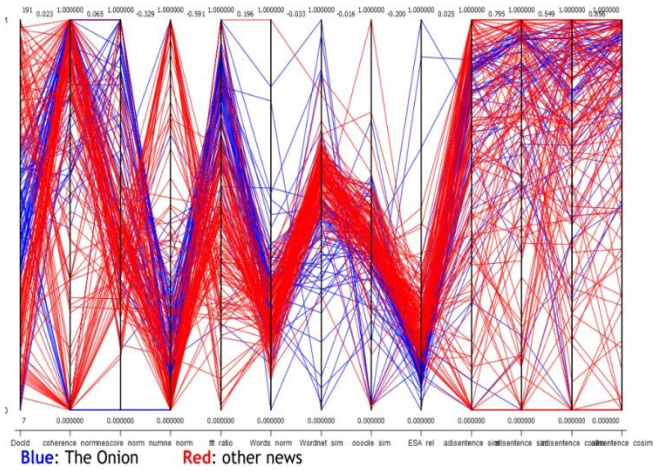


Figure 4. Orange data analysis workflow.

Another interesting result was provided by the investigation of semantic similarities: while ESA proved to have no influence, Wordnet Similarity proved to be the best evidence for creative texts, showing the fact that the concepts are highly connected. This fact gives credit to the definition of creativity in the sense that the more semantic similarity exists between the words, the better qualitative the text is. From the four different options for LSA, the one that proved to be the most correlated with creativity was the average cosine similarity for all sentences. High values for this measure witnessed for non-creative texts, which is natural since LSA reflects the words connections that could be seen in the training corpus (showing a higher "usualness" of the text than the word pairs with smaller LSA scores).

Finally, the analysis of Coherence Measure did not bring anything new, proving that no matter how creative a text is, it should be coherent.

The obtained model was tested in a 10-fold cross validation setup, starting from the assumption that the news taken from the The Onion were creative, while the others were not. The results are presented in Table 1:

TABLE I. EXPERIMENT RESULTS

Values prediction				Confusion matrix	
Real / Predicted	Creative	Non-creative	Sum	Precision	Recall
Creative	46	15	61	0.754	0.6866
Non-creative	21	103	124	0.8306	0.8729
All	67	118	185		

The accuracy for this experiment was 80.54%, which is quite high, considering the difficulty of this task.

B. *Adaptability Experiment*

In order to evaluate the adaptability capacity of our classifier, we tried to evaluate a different type of texts (book reviews taken from [19]) using the same classifier as for the Assessment Experiment. Therefore three masters' students individually evaluated 20 different book reviews, assessing

to each of them a rank between 1 and 3. One was assigned to creative texts, two was assigned to mildly creative texts, while three was assigned to non-creative texts (see Table 2). Unfortunately, the inter-rater agreement Kappa Statistic [20][21] was low (perceived agreement was Po = 0.45), which according to the Kappa interpretation done by Altman [22], was not enough to further consider this ranking. Therefore, in order to improve this situation, we considered instead a binary classification. In order to decide what to do with the mildly creative texts (the ones evaluated with the rank 2), we tested two different situations (evaluating these problematic documents in both possible ways). In the first one, we considered that they were creative, so we evaluated the text as being creative if they formerly received the rank 1 or 2, and non-creative if they received rank 3. Here, the value for the Kappa Statistic was Po = 0.633. In the second situation, we considered that only the texts evaluated with rank 1 were creative and the rest were classified as non-creative (see Table 3). This time, the Kappa Statistic was Po = 0.733. The higher Kappa Statistic score from the second situation gave us a hint that this should be the correct binary classification of the reviews. This decision was also enforced by the fact that, using the majority class (creative/non-creative) amongst the reviewers as the gold standard, in the first situation we ended up with 12 creative texts (out of 20), while in the second we had only 4 texts that were considered to be creative. Since our hypothesis was that there are more non-creative texts than creative ones, the second decision augments the decision made starting from the inter-rater agreement.

After deciding how to consider the reviews initially evaluated with rank 2 and computing the inter-rater agreement, we tried to correlate the output provided by the previously built model with the classes obtained from the reviewers' gold standard evaluations.

Unfortunately, all the reviews were classified as being non-creative, missing 4 creative texts – R5, R6, R9 and R13 – (see Table 3). This might be due to the fact that the built classifier is too specific for The Onion news, and does not find book reviews as creative. However, it should be noted that from these four misclassified reviews, only one was considered creative by all three reviewers. The experiment's accuracy was 80%.

TABLE II. THE RANKS PROVIDED BY THE 2 REVIEWERS FOR THE 20 BOOK REVIEWS CONSIDERING A SCALE WITH 3 VALUES:1 FOR CREATIVE, 2 FOR MILDLY CREATIVE AND 3 FOR NON-CREATIVE

Filename / Reviewer	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12	R 13	R 14	R 15	R 16	R 17	R 18	R 19	R 20
Reviewer1	3	3	2	2	2	2	1	3	2	2	2	2	1	3	3	3	2	1	3	3
Reviewer2	3	3	2	3	1	1	3	3	1	1	2	3	1	1	3	1	2	3	3	2
Reviewer3	2	3	3	2	1	1	3	3	1	2	3	3	1	2	3	2	2	2	2	3

TABLE III. THE FINAL EVALUATION OF THE 20 BOOK REVIEWS

Filename / Evaluation	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12	R 13	R 14	R 15	R 16	R 17	R 18	R 19	R 20
Non-creative	3	3	3	3	1	1	2	3	1	2	3	3	0	2	3	2	3	2	3	3
Creative	0	0	0	0	2	2	1	0	2	1	0	0	3	1	0	1	0	1	0	0

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a model that was intended to discriminate creative from non-creative news articles that was built combining nine different measures. The model could be improved by removing or changing the important assumptions done during the course of this work, such as the The Onion articles being always creative (while the news articles are not) and using just a binary creativity scale: creative and non-creative.

The first part of the experiment presented in this paper delivered the following specific conclusions:

- Word Norms Fraction was the measure that was best correlated with the lack of creativity, which was expected considering the definitions of creativity and of Word Norms Fraction. Google Similarity Distance was in the same situation;
- Named Entities analysis showed that they are signs of a creative text as long as not too many distinct such entities are used;
- Wordnet Similarity proved to be the best evidence for creative texts, while LSA was similar to the measures of Word Norms Fraction and Google Similarity Distance in providing a measure for text "usualness" and therefore giving evidence of non-creative texts. They also have similar weights in the final classifier. ESA had no influence in the built classifier;
- Less coherent texts were expected to be more creative but coherence score was found to have no influence in identifying creativity.

The second part of our experiment investigated the possibility of generalizing the built classifier so that it can be applied to different kinds of texts and/or topics. The difficulty of this task was observed in the very low inter-rater agreement – we believe that more judges are needed to obtain better agreement and build a more robust data set. Also, a finer scale would be useful to cope with the problematic of "how creative" means creative, and to give a better idea of creativity than plain binary values.

 ISBN: 978-1-61208-280-6

The fact that the model built during the first experiment did not consider any review as being creative might be due to the fact that it is tested on a different corpus. It seems also that there are "levels" of creativity according to the analyzed texts: a satire news articles domain may be more creative than books reviews, in general. Thus a bigger data set, comprising different text sources, may achieve better results.

Even though the classifier did not detect creative reviews, the results of both experiments were around 80%, showing that there might be a possibility that the classifier adapts well to different domains and kinds of texts. However, the lack of true positive examples from the second experiment makes us be a little cautious in clearly stating this fact.

These results made us question whether we really identified creativity or we identified a solution to another very difficult problem: satire detection in texts.

The classifier performed reasonably well at differentiating articles from The Onion and from other serious news websites. We believe that increasing the size of the data set, and testing it further, could confirm our assumption. It also shows that satire and creativity are related, since we were searching for creativity but we may have ended up in identifying satire. Previous work has been done about satire detection [23], but increasing the emphasis on semantic similarity, as this work does, could yield better results than those in the referred experiment.

As future work, we plan to verify our assumption related to what make the The Onion articles special (are they expressing creativity, satire, or have we made a wrong assumption considering them to be special?). We intend to do this by using manually classified texts to train the model and then to use it in order to decide whether any of the two assumptions stands and which of them is more adequate.

ACKNOWLEDGMENT

The research presented in this paper was supported by project No. 264207, ERRIC-Empowering Romanian Research on Intelligent Information Technologies/FP7-REGPOT-2010-1.

We would like to thank to the three students that helped us in our experiment (Rajani, Agata and Pavel) and to the reviewers that pointed out an important error that was present in the initial submission. Fixing it greatly improved our results and (we believe that) the paper looks much better now.

REFERENCES

[1] X. Zhu, Z. Xu, and T. Khot, "How creative is your writing? a linguistic creativity measure from computer science and cognitive psychology perspectives," in Proceedings of the Workshop on Computational Approaches to Linguistic Creativity, ACL, 2009, pp. 87-93.

[2] X. Hu and B. Yu, "Exploring the relationship between mood and creativity in rock lyrics," 12th International Society for Music Information Retrieval Conference (ISMIR 2011), 2011, pp. 789-794.

[3] R.L. Cilibrasi and P.M.B. Vitanyi, "The google similarity distance," IEEE Transactions on Knowledge and Data Engineering, 19(3), 2007, pp. 370-383.

[4] E. Gabrilovich and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," Journal of Artificial Intelligence Research, 34(2), 2009, pp. 443-498.

[5] About WordNet - WordNet - About Wordnet [online] http://wordnet.princeton.edu/ [retrieved; May, 8, 2013]

[6] J. He, W. Weerkamp, M. Larson, and M. de Rijke, "An effective coherence measure to determine topical consistency in user-generated content," International journal on document analysis and recognition, 12(3), 2009, pp 185-203.

[7] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, 41(6), 1990, pp. 391-407.

[8] A. Renouf, "Tracing lexical productivity and creativity in the british media: the chavs and the chav-nots," in Lexical Creativity, Texts and Contexts, John Benjamins Publishing Company, Amsterdam, 2007, pp. 61-89.

[9] T. Veale, "Creative language retrieval: A robust hybrid of information retrieval and linguistic creativity," in Proceedings of ACL 2011, 2011, pp. 278-287.

[10] A. Jordanous, "A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative," Cognitive Computation, 4, 2012, pp. 246-279, ISSN 1866-9956.

[11] Z. Kovecses, "A new look at metaphorical creativity in cognitive linguistics," in Cognitive Linguistics 21(4), 2010, pp. 663-697.

[12] T. Veale, "An analogy-oriented type hierarchy for linguistic creativity," in Knowledge-Based Systems, 19, 2006, pp. 471-479.

[13] A. Lehrer, "Blendalicious," in Munat, J. (Ed.) Lexical creativity, texts and contexts. John Benjamins, Amsterdam, 2007, 115-136.

[14] USF Free Association Norms: Introduction [online] http://web.usf.edu/FreeAssociation/Intro.html [retrieved; May, 8, 2013]

[15] Total Number of Pages Indexed by Google | Statistic Brain [online] http://www.statisticbrain.com/total-number-of-pages-indexed-by-google/ [retrieved; May, 8, 2013]

[16] The Onion - America's Finest News Source [online] http://www.theonion.com/ [retrieved; May, 8, 2013]

[17] Orange – Data Mining Fruitful & Fun [online] http://orange.biolab.si/ [retrieved; May, 8, 2013]

[18] TANAGRA - A free DATA MINING software for teaching and research [online] http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html [retrieved; May, 8, 2013]

[19] Maite Taboada [online] http://www.sfu.ca/~mtaboada/research/SFU_Review_Corpus.html [retrieved; May, 8, 2013]

[20] J. Cohen, "A Coefficient of Agreement for Nominal Scales" in Educational and Psychological Measurement 20, 1960, pp. 37-46, DOI: 10.1177/001316446002000104.

[21] J. Carletta, "Assessing agreement on classiffication tasks: the kappa statistic," Computational Linguistics, 22(2), 1996, pp. 249-254, ISSN 0891-2017.

[22] D.G. Altman, Practical Statistics For Medical Research. Chapman & Hall, 1991, ISBN 9780412276309.

[23] C. Burfoot and T. Baldwin, "Automatic satire detection: are you having a laugh?," in Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, ACLShort '09, 2009, pp. 161-164.

 ISBN: 978-1-61208-280-6

StreamQuilt: A Timeline-Aware Integration of Heterogeneous Web Streams

Riho Nakano
Graduate School of Media and Governance
Keio University
Fujisawa, Kanagawa, Japan
e-mail: nrihos@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 timeline-aware integration system for web streams, such as micro-blogs and real-time information. This system, called StreamQuilt, analyzes implicit and temporal context-dependent relationship among heterogeneous media streams on the web. Our concept is to capture features from a media stream according to its temporal status and relationships to other streams. This system assigns different features to the stream at the different time. This system provides a synchronized streams association mechanism generate a new stream by evaluating the implicit relevance between heterogeneous streams along a timeline. This mechanism utilizes a sequence of data filters to be applied to different filters to streams. Each filter removes contextual ambiguity and various noises from media streams. This approach is advantageous in providing new information by detecting implicit relationships, such as cause-effect relationship and provider-consumer relationship, among web streams. Our system is applicable to mobile advertisement, participatory entertainment systems, and sentiment analysis of social networking services.**

Keywords-Cross-media infrastructure; heterogeneous stream association; implicit relationship analysis.

I. INTRODUCTION

Streams, such as television streams and radio streams, are among the most important media types in our daily lives. Although the number of streams in era of television and radio was hundreds at most, the popularization of broadband networks and high-performance devices has provided us with opportunities to access millions of streams nowadays via the Internet. Such rapid popularization of streams has generated a serious demand for extracting useful information and inter-stream relationships by conducting an integrative analysis of streams.

However, there are no services that integrate heterogeneous streaming media, such as television, radio, live Internet video, and live messages on social network services (SNSs). In order to compute the implicit relationships among those streaming media, it is necessary to deal with temporal changes in the contexts of the streams. To develop a novel stream management system that integrates heterogeneous streams and analyzes the integrated streams to extract useful information, two factors should be considered. The first factor is the heterogeneity in data structures of the streams. Because each stream has its own story and organization of data, it is difficult to make direct comparisons. The second factor is temporal-context dependency in relationships between streams. The meaning and position of each element are difficult to identify uniquely, because the meanings of and inter-relationships between elements in a stream should be interpreted according to its own temporal context. It is essentially a novel stream management method to integrate everyday streaming media data.

Toward the above objectives, we propose in this paper a media stream management system for both Internet media and conventional broadcast media such as television and radio. This system, called "*StreamQuilt*", provides a dynamic integration method for heterogeneous media streams. The key technology of this system is a data analysis functions for detecting implicit and temporal-context-dependent relationships between streams by removing contextual ambiguity and various noises from media streams. This method invokes a data analysis filter associated with a specific time window in order to extract a temporal-context dependent metadata from a media stream. The system then generates comparable feature sets from diverse types of data in order to find the highly inter-related fragments in different streams by evaluating correlations between metadata generated from those heterogeneous streams.

The advantage of the system is a new data provision for detecting implicit relationships, such as cause-effect relationships and provider-consumer relationships, between heterogeneous streams. Our system configures a metric space by removing irrelevant features to calculate relevance score of streams at a specific time. Each metric space consists of features related to the specific time content. This mechanism is effective for calculating the context-dependent relevance score of elements from streams. According to this context-dependent relevance score, our system integrates heterogeneous streams to create a new data stream consisting of contextually related information. This system is applicable to mobile advertisements, participatory entertainment systems, sentiment analysis on SNSs.

As an example of such implicit relationships between streams, inductive relationships are found between a television stream and an SNS stream. In this case, content broadcast on television causes various reactions in SNS communications. The system extracts such indirect effects of one stream on the other stream by triggering filters for the SNS according to information in the television program.

 ISBN: 978-1-61208-280-6

Figure 1. Screenshot of Community Generation System

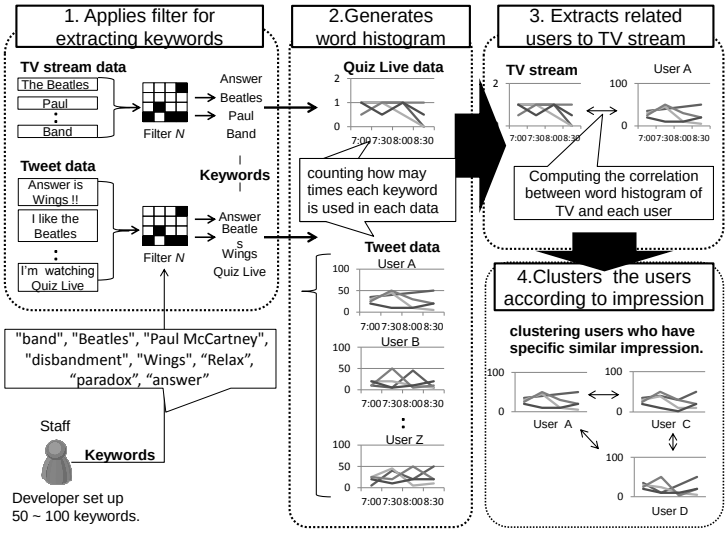


Figure 2. Concept of Community Generation System based StreamQuilt system

The information obtained about implicit relationships is very useful to plan a new television show and tie-in events on SNSs.

The remainder of this paper is structured as follows. Section II describes a temporal-context analysis of the relationship between a television stream and an SNS stream as a motivating example. Section III discusses several related studies. Section IV shows an architectural overview of our system. Section V shows the prototype of our system. Section VI evaluates the effectiveness of our system. Finally, Section VII concludes this paper.

II. MOTIVATING EXAMPLE

In this section, we present a motivating example of our StreamQuilt system. For example, there are staffs of TV program "Quiz LIVE", who want to edit program according to reactions of TV viewers during broadcasting. To get live feedback from viewers, the staffs tried to use a community generation system [1] for their Quiz LIVE. This system automatically groups viewers who have similar impression about "Quiz LIVE" as community, by analyzing SNS user messages. To realize the community generation system, the staffs can use our StreamQuilt. In this case, our StreamQuilt system extracts similarities between SNS users according to topic of TV program. This system is effective for two official positions in television station follow as.

1) Staffs who produce TV program in studio: The staffs can know name or number or member of community, through inputting Quiz LIVE data. From the information, they can judge how viewers feel, and how many viewers do so. Here, before the staffs broadcast the program, they need to set up 50~100 keyword sets to the system in each topic of Quiz LIVE. The keywords are noun or adjective for impressions. For example, there is a question "What is name of the band group that Paul McCartney formed after the Beatles disbanded?" in the program. Also, the question has three options as the answer: A)"Wings", B)"Relax", C)"Paradox". For this topic, they input "band", "Beatles", "Paul McCartney", "disbandment", "Wings", "Relax", "paradox", "answer". Figure 1 illustrates a screenshot of the community generation system. When the system starts with a broadcasting of Quiz LIVE, it visualizes community information on a view space according to the set topic in the program. Also, the system dynamically displays the program rating on twitter with using a graph. In addition, the staffs can see a transition of the information from the past to the present by dragging slider (Figure 1). While the staffs are broadcasting the program, they try to exchange the next on-air question according to interests or thoughts of viewers by checking community information.

2) Staffs who develop community generation system: When the staffs implement the community generation system, they need to suppose text data of two types as stream data: A) Quiz LIVE data is text as an annotation of the program (eg., talent, title, staff, topic of television contents). B) tweet data consists of its message and its time information and user id. The staffs design the system as follows (Figure 2). First, the system continuously reads Quiz LIVE data and tweet data. Secondly, from these data, the system extracts keywords "band", "Beatles", "Paul McCartney", "disbandment", "Wings" etc. And then, the system generates a word histogram of Quiz LIVE and each user by checking how frequently each keyword is used in Quiz LIVE data and their tweet data. Next, the system extracts users who tweeted about Quiz LIVE by computing the relevance of the word histogram of Quiz LIVE data and their tweet data. In the same way, the system is able to extract user who thinks that Paul McCartney formed "Wings" by comparing the word histogram of "Paul McCartney", "Wings" and users. In addition, the system is able to group users into several communities automatically

 ISBN: 978-1-61208-280-6

who love Beatles by computing the similarities of their word histogram.

III. RELATED WORK

In this section, we summarize the related work of our approach. Our approach focuses on the integration of heterogeneous media streams by considering their temporal-context-dependent relevance. This approach is related to the traditional data mining methods such as [2][3][4]. Those data mining methods focus on finding valuable knowledge from the large-scale homogeneous data such as relational tables. There are several studies [3][4] that apply such data mining concepts and techniques into social network data. For example, the system in [3] recommends users to each other by clustering users into several groups. The system presented in [4] finds semantic similarities between users by comparing SNS user profiles. Those approaches are effective to integrate large-scale profiles and SNS activities. However, they are not suitable to integrate heterogeneous media streams derived from various network resources such as TV, SNS, and other streaming services.

As methods for extending the conventional data mining technologies to support heterogeneous social network analysis, a behavior targeting (BT) advertisement [5] is a widely adopted approach to recommend user an appropriate advertisement by analyzing the stored user behavior data. The data analysis methods proposed in [6][7] extract a user preference by applying machine learning method and data mining methods to a user search history and/or access log to web site. And then, they compute the relevance of online-advertisement and the user preference. BT is effective to find the relevant item data by computing the relation of stored stream data and the item data. Those approaches merely use timeline of data streams because those approach focuses on finding a valuable patterns or knowledge commonly appearing on data streams.

These are several methods to find a change of relations or patterns for differing from the conventional data mining methods to find them at all time. For example, these system[8][9][10] find most frequently used words and topics and their volume of sentiments such as "positive" or "negative", from tweet on twitter according to time. These approaches are effective for detecting trend of social activity in a specific SNS by classifying similar data in chronological log of posts. These approaches are not suitable to compute the similarities among stream data in heterogeneous media.

Our StreamQuilt system differs from these existing data mining methods in the following two aspects. First, our system indirectly computes the relevance of heterogeneous stream for various type data. Here, "indirectly" means that the system compares feature values extracted from media data and doesn't compare the media data by itself in order to compute the relevance. Second, the system converts streams into feature values in same temporal range because the system analyzes multi streams which share same timeline. Concretely, the system sets particular feature sets to filter according to time window. The system extracts feature value from different streams by using same feature sets of the filter.

As an approach similar to ours, there are studies [11] [12] for clustering TV viewers on SNS or for extracting short messages related to TV from SNS. For example, Doughty et al [11] analyze audience networks by detecting a connection between users through communication about a specific TV program. A filtering method proposed in [12] computes similarities between feature keywords of Twitter message and particular program in order to collect relevant tweet. These approaches are effective for comparing constant features of TV program and tweets without timeline, by extracting features in particular time or one TV program. They are not suitable to find similarities between program and tweets in variation of feature. Our approach extracts long-term changes of features values from each stream by reducing values of irrelevant feature along with timeline. The long-term changes of features value indicate how one stream affects other stream not only immediately but also previously or subsequently. Our system is able to find stream data, which has the changes of feature values similar to other stream with them.

IV. SYSTEM ARCHITECURE

Figure 3 illustrates the architectural overview of the proposed system. StreamQuilt system consists of the three main components as follows: 1) an event-driven engine, 2) a filter module and 3) a relevance computation module. The event-driven engine invokes the filter module when the module recognizes matching of the current time and property of filters. And then, the module sends the matched filter into the filter module. The filter module generates metadata of streams by using the filter. Concretely, this module extracts specific data from streams as metadata according to feature set of the filter. In addition, this filter module sends the metadata to the relevance computation module. The relevance computation module calculates the correlation among the metadata sent by the filter module.

The most important mechanism of our system is to switch filter according to time. Our system has diverse filters corresponding to a specific time window because the system needs to extract changing different feature of stream according to time. The filters specify feature keyword as a criterion of similarity among streams. To realize this mechanism, the system provides the event-driven engine. This event-driven engine enables the filter module to extract different metadata from stream data at each time window. By using the proper metadata, the relevance computation module is able to output changing relevance of streams with time.

4.1 Data Structure

The data structure in this system consists of four data which are now explained in detail as follows.

 ISBN: 978-1-61208-280-6

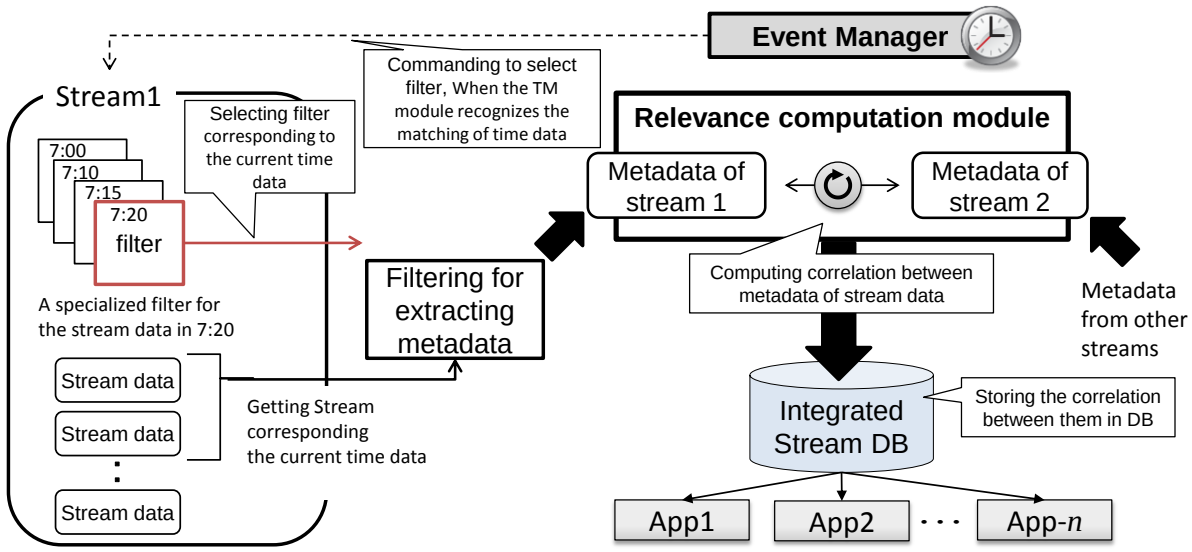


Figure 3. Architectural Overview of StreamQuilt Having an Event-driven Engine to Drive Heterogeneous Stream Integration

Snippet data s is a primary data structure in stream. The snippet data s is modeled as follows: $s := \{t, contents\}$, where s is a tuple consisting of timestamp information denoted by t and $contents$ (eg., text, image, moving image). The timestamp information t in snippet data s represents a time when the snippet data s flows into the stream.

Filter is a data for designating time window and feature set. This filter μ is modeled as follows: $\mu := \{t_s, t_e, f\}$, where t_s and t_e, which are start and end of time, are temporal range corresponding to the filter. f is feature set consisting of feature keyword and weight pairs. The weight is strength of the feature keyword in stream data. f is modeled as follows: $f := \{< k_1, w_1 >, < k_2, w_2 >, \dots, < k_n, w_n >\}$, where $k_{[1 \dots n]}$ is a feature keyword and corresponds $w_{[1 \dots n]}$.

Metadata is feature set consisting of feature keyword and value pairs. The feature keyword is the same as feature keyword of filter. A metadata a is modeled as follows: $a := \{< k_1, v_1 >, < k_2, v_2 >, \dots, < k_n, v_n >\}$, where $k_{[1 \dots n]}$ is a feature keyword and corresponds $v_{[1 \dots n]}$.

4.2 Functions

StreamQuilt system consists of two core function as follows; A) a metadata generation function and B) a relevance computation function.

Metadata generation function selects set of snippet data according to the current time in order to convert the set of snippet data into metadata. This function accepts the set of snippet data and filter and returns metadata. Metadata generation function is modeled as the equation (1).

$$f_{mgneration}(\langle s_1, s_2, \cdots s_i, s_{i+1} \rangle, \mu_j) \rightarrow a \quad (1)$$

where s_i denotes a i-th data and μ_j denotes a j-th filter. Here, The function accepted i+1 snippet data. This function computes a frequency or ratio of each feature keyword $k_{[1 \dots n]}$ in the set of snippet data. Concretely, this function counts how many times it gets m-th feature keyword from contents of set of snippet data. In addition, the function multiplies the count of m-th feature keyword and m-th weight $w_{[m]}$ in filter. This function sets the m-th product to m-th value of metadata.

Relevance computation function calculates the inner product between metadata a and a'. The function is defined as the equation (2).

$$f_{revcomputation}(a, a') := \sum_{m=0}^{n} a_{[m]} \cdot a'_{[m]} \quad (2)$$

where the top- n is the number of feature keywords in a and a', and m is the m-th feature keyword.

V. IMPLEMENTATION

In this section, we implement a prototype of the StreamQuilt system. In this implementation, our system dynamically visualizes correlations between a specific tweet stream and video according to time to play video. Figure 4 shows the detailed architecture of our prototype system. On server side, the system processes from collecting tweet data to computing the relevance scores. On client side, the system visualizes the relevance score to end-user.

The server side module consists of the following five components: stream data collector, database management system, timeline management engine, feature conversion system, relevance computation. We describe these five components in detail. The stream data collector gathers tweet data at regular intervals by twitter API. Timeline management engine gets time data to play from video elements. When the engine recognizes matching of the time to play and time scope of video filters, it starts the feature

 ISBN: 978-1-61208-280-6

conversion system. The feature conversion system consists of the following sub-components; filter selector and feature extraction engine. The filter selector switches the filters according to time scope to apply of the filters. The system computes words frequency in tweets by the selected filter in order to extract feature metadata. Also, the system uses the filter as feature metadata of video. The relevance computation calculates the correlation score between feature metadata of tweets and video. The client side program dynamically re-visualizes the correlation with graph on web browser by utilizing HTML5 and jQuery API.

The most important feature of this implementation is compatibility between it and web technology for getting tweet data and for processing them. This is because we can get stream data through existing API. Many useful API are provided by SNS such as Twitter and Facebook. In addition, we need a technology which enables to process stream data effectively in real time. So, we develop the prototype system by Node.js which is the modern server side web technology with JavaScript. Node.js uses an event-driven, non-blocking I/O model that makes it lightweight and efficient for data-intensive real-time applications that run across distributed devices. By utilizing Node.js, the system is able to sequentially process a large amount of requests from an engine that starts an analyzing process. It is effective for the system to permit other processing to continue, before inputting or outputting of data has finished by utilizing non-blocking I/O.

VI. EVALUATION

5.1 Outline of Experimental Studies

In this section, we evaluate the effectiveness of our system when applied to TV and Twitter. For this, we implement the Stream Quilt prototype system. This evaluation experiment investigates the timeline management to filter out noisy data and compute the relevance score in each time. The experiment evaluates the precision of relevance score. Here, the precision means the number of irrelevant tweet reduced by filter in each time window. We compare our relevance computation with time manager to conventional relevance computation without time manager. We have used filters for each relevance computation: A) feature sets cover any time window; and B) the feature sets are special for each time window and contain no irrelevant features. We show that to change the feature sets makes a significant contribution to the relevance computation for stream data.

For this experiment, we configured tweets and a television program as stream data. We chose a famous music program called “Kouhaku” which is broadcast once a year. This is because we can get more tweet data about Kouhaku than tweets about other television program. In Kouhaku, each artist appeared for about 5 min. So, we gathered 1) a random set of 500 tweets by searching for the hashtag “#Kouhaku” every 5 min, 2) one feature set for

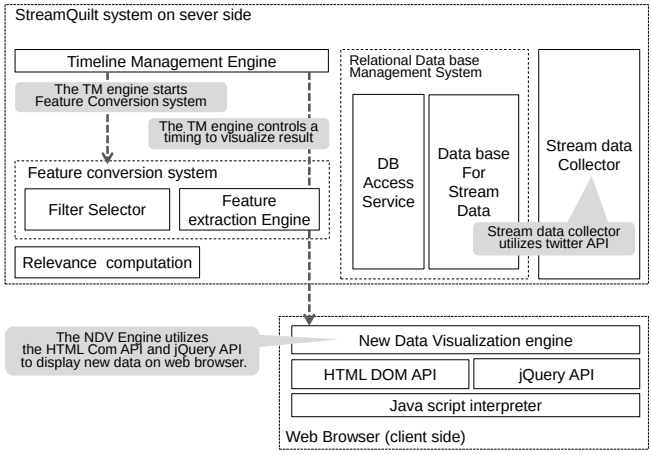


Figure 4. Prototype System Architecture of the StreamQuilt System Implemented Using Modern Technologies.

Method A, and 3) thirteen feature sets for Method B. These feature sets were formed in recognition of the basic structures of our filters. In all, we used fifteen data filters and 6500 tweets for 65 min. We employed words of nouns or adjective as feature sets for Kouhaku. We defined keywords and the values (weight) in various situations (TABLE I DEFINITION OF FEATURE SETSTABLE I). More specifically, the feature sets in method B consist mainly of three types as follow; a) keywords such as "Arashi" or "MC" appear through one hour. b) keywords such as "Opening" or "Desney" appear on only particular time. c) keywords such as "Top batter" or "character" appear according to keywords of type b, and they are related to keywords of type b. Values of these types change through one hour as in Figure 5.

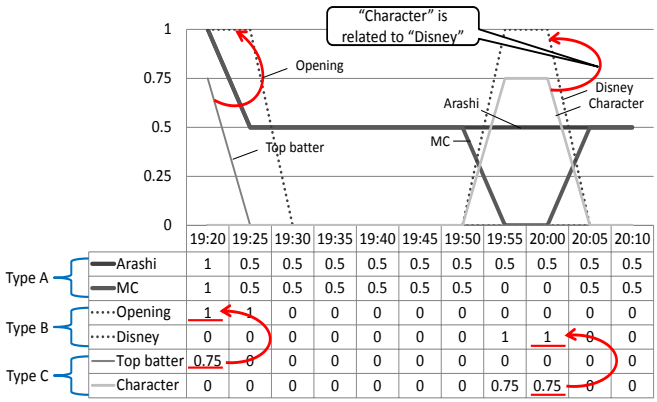
TABLE I DEFINITION OF FEATURE SETS

Value	Situation to use keyword	Example keywords
0.5	Any topic in “Kouhaku”	MC, team
0.75	Explanation of background for attention to any topic in “Kouhaku”	singer, top batter, character, Fukuoka (singer’s home)
1.0	A specific topic in “Kouhaku”	Opening, Disney

5.2 Experimental Result and analysis

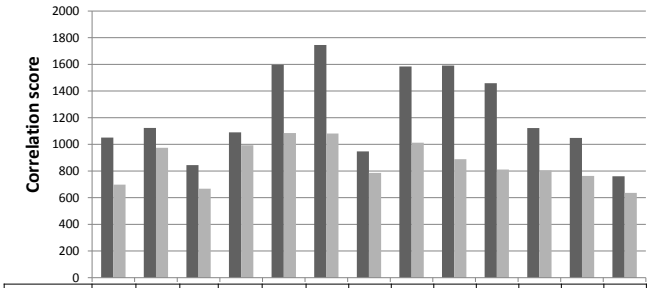
In this section, we evaluated the relevance computation in the previous subsection in order to clarify the effectiveness of our approach. Figure 6 shows the correlation between the program Kouhaku and the Twitter hashtag #Kouhaku for each approach. This result shows that score by Method B (using different feature sets at each time) is less than score by Method A (using the same feature sets at all time). It means that the relevance computation by Method B performs better than that by Method A. This is because Method B recognized more noisy tweets than Method A. Although the tweets obtained from the hashtag Kouhaku over the period are largely related to the program Kouhaku, several tweets in any specific topic are unrelated tweets. For

 ISBN: 978-1-61208-280-6



		19:20	19:25	19:30	19:35	19:40	19:45	19:50	19:55	20:00	20:05	20:10
Type A	Arashi	1	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
	MC	1	0.5	0.5	0.5	0.5	0.5	0.5	0	0	0.5	0.5
Type B	Opening	1	1	0	0	0	0	0	0	0	0	0
	Disney	0	0	0	0	0	0	0	1	1	0	0
Type C	Top batter	0.75	0	0	0	0	0	0	0	0	0	0
	Character	0	0	0	0	0	0	0	0.75	0.75	0	0

Figure 5. three types of feature sets in Method B



Time	7:20	7:25	7:30	7:35	7:40	7:45	7:50	7:55	8:00	8:05	8:10	8:15	8:20
Score By Method A	1051	1123	844	1090	1589	1745	947	1584	1591	1458	1122	1048	760
Score By Method B	698	974	667	993	1084	786	1012	889	889	822	804	763	636

Figure 6. Correlation between "Kouhaku" and the Twitter "hashtag Kouhaku"

example, many users tweeted about the artist "Nana Mizuki" even though that the artist did not appear in 7:25. Hence, the system needs to recognize tweets about Nana Mizuki as noisy data because those are unrelated to the topic in 7:25. This shows that our system effectively reduces noise for relevance computation by setting relevant feature sets to filter in specific time windows.

VII. CONCLUSION

We have proposed a timeline-aware integration system for web streams, is called "StreamQuilt" that evaluates the implicit relationship among heterogeneous streams. This system extracts the indirect relevance between streams by calculating correlation among comparable metadata. Here, the metadata are extracted from various data type in streams. This system uses the relevance score for creating a new data stream consisting of the contextually relevant information. The unique feature of this system is a timeline-based data analysis mechanism that applies different filter in each time window. This mechanism generated a specific metric space by using feature sets related to the time window. This mechanism is able to reduce sequentially noisy stream data by computing the distance between stream data in the metric space generated.

ACKNOWLEDGMENTS

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] R. Nakano, and S. Kurabayashi, "A Stream-Oriented Community Generation for Integrating TV and Social Network Services", The Seventh International Conference on Internet and Web Applications and Services, May. 2012, pp.286-289.

[2] W. Wu, and L. Gruenwald, "Research Issues in Mining Multiple Data Streams" Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques – StreamKDD, 2010, pp. 56-60.

[3] A. Voulodimos, C. Patrikakis, P. Karamolegkos, A. Doulamis, and E. Sardis, "Employing clustering algorithms to create user groups for personalized context aware services provision" Proceedings of the 2011 ACM workshop on Social and behavioural networked media access - SBNMA '11, 2011, pp. 33-38.

[4] C. Akcora, B. Carminati, and E. Ferrari, "Network and profile based measures for user similarities on social networks" 2011 IEEE International Conference on Information Reuse & Integration, August. 2011, pp. 292-298.

[5] D, Anjali, G. Amar, and V. Samjeev, "Data Mining Techniques for Optimizing Inventories for Electronic Commerce" Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00, 2000, pp. 480-486.

[6] G. Xiao, Z. Gong, and J. Guo, "Personalized Scheduling Search Advertisement by Mining the History Behaviours of Users" 2009 IEEE International Conference on e-Business Engineering, October. 2009, pp. 29-36.

[7] D. Hu, Q. Yang, and Y. Li, "An algorithm for analyzing personalized online commercial intention" Proceedings of the 2nd International Workshop on Data Mining and Audience Intelligence for Advertising - ADKDD '08, 2008, pp. 27-36.

[8] K. Vinh, S. Chaitanya, R. Rajiv, and R. Jay, "Towards Building Large-Scale Distributed Systems for Twitter Sentiment Analysis" Proceedings of the 27th Annual ACM Symposium on Applied Computing - SAC '12, March. 2012, pp. 459-464.

[9] H. Wang, D. Can, A. Kazemzadeh, B. François, and N. Shrikanth, "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle" ACL '12 Proceedings of the ACL 2012 System Demonstrations, July. 2012, pp. 115-120.

[10] A. Marcus, M. Bernstein, O. Badar, D. Karger, S. Madden, and R. Miller, "TwitInfo: Aggregating and Visualizing Microblogsfor Event Exploration" Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11, May. 2011, pp. 227-236.

[11] M. Doughty, D. Rowland, and S. Lawson, "Who is on Your Sofa? TV Audience Communities and Second Screening Social Networks", Proceedings of the 10th European conference on Interactive tv and video - EuroiTV '12, 2012, pp. 79-86.

[12] D. Dan, J. Feng, and B. Davison, "Filtering microblogging messages for social tv", Proceedings of the 20th international conference companion on World wide web - WWW '11, 2011, pp. 197-200.

 ISBN: 978-1-61208-280-6

Designing Formulas for Creating a Healthcare Cloud Service Based on a Formula Calculation Platform Service

Xing Chen
Department of Information & Computer Science
Kanagawa Institute of Technology
Atsugi-shi, Kanagawa, Japan
chen@ic.kanagawa-it.ac.jp

Keiichi Shiohara
Department of Information & Computer Science
Kanagawa Institute of Technology
Atsugi-shi, Kanagawa, Japan
s1385006@cce.kanagawa-it.ac.jp

Hikaru Tazumi
Department of Information & Computer Science
Kanagawa Institute of Technology
Atsugi-shi, Kanagawa, Japan
s1021101@cce.kanagawa-it.ac.jp

***Abstract*— Cloud services provided for healthcare are easy to use and updated quickly with new knowledge therefore they are widely utilized in nowadays. A research issue here we need to solve is to provide a no-programming method for developing cloud services. The research issue comes from the background that the specialists in the area of healthcare are not also the specialist in the area of information technology. In this paper, we propose a method for creating a healthcare cloud service based on a cloud service that we developed referred to as the Formula Calculation Platform Service. The important feature of the method is that service builders only need to consider about the functions of their services. Based on the proposed method, we created a cloud service for weight control to evaluate the effectiveness of the proposed method. The effectiveness of the proposed method is confirmed by demonstration experiments.**

Keywords-cloud service; programming; platform; healthcare; weight control.

I. INTRODUCTION

Healthcare services, such as businesses-oriented health management system [1][2] and general health support services for individuals and families [3], etc., are highlighted gradually. A wide variety of services for healthcare are developed, which are utilized regardless of gender and age, etc. In order to let the services to be widely utilized, it is important to update the service constantly based on the new knowledge. When users utilize a service for healthcare at first time, they do not know whether the service suits them or not. Therefore, it is a common case that they first try it to see if it suits them or not. Depending on the needs of the users, it is important to update services continuously or to add new functions to prevent users boring the services and help spread the services to be widely utilized.

We focus our research on methods of building cloud services. We have developed a cloud service platform, referred to as a Formula Calculation Platform Service (FOCAPLAS) [4]. By using this service, without programming, cloud services can be implemented based on mathematic formulas uploaded by service providers. In the same way, using this service, cloud services can also be updated when mathematic formulas are modified.

In this paper, we represent formulas for implementing a weight control cloud service as an example showing the efficiency of FOCAPLAS. Based on the proposed formulas, a weight control cloud service is created without program coding. We also performed demonstration experiments for the created cloud service. Our experimental results made it clear that data of the service users are correctly stored, formulas, which are uploaded by service providers, are executed correctly. Therefore, we confirmed the validity of our proposed method.

The rest of the paper is structured as follows. Section II presents related researches in this research area. Section III presents the outline of FOCAPLAS. Section IV explains the basic idea of constructing the cloud service for the healthcare. After this, Section V presents an implantation example for developing a weight control cloud service. Our experimental results are also presented in this section. Technical details about communication protocols are presented in Section VI. Finally, Section VII concludes this paper.

II. RELATED RESEARCHES

The purpose of healthcare services is to provide healthcare related comments, or to provide healthcare related computing results based on the data of the users. From technical point of view, healthcare services can be considered as a kind of knowledge processing system.

In the research field of Information and Communication Technology (ICT), knowledge processing is an important research field. Various methods related to the knowledge processing have been proposed [5]. In particular, as the developing of ICT, a large amount of knowledge information data is distributed in cloud, research studies on knowledge information retrieval [6], and knowledge information coordination [7] are performed. In addition, the research

 ISBN: 978-1-61208-280-6

studies are also performed on providing knowledge services [8], and knowledge sharing and recommendations [9] based on the stored data of knowledge information in servers.

As not only knowledge contents, but the calculation results are also required to be provided, services are developed, which provide the calculation results based on knowledge [10][11].

In the case that automatic calculation based on the knowledge is provided as a service, it is required to code a program and build a server to run the program. Traditionally, in order to build the server, high-ICT knowledge and high-quality computers are required. Currently, as pre-built server platforms are provided as cloud services [12][13], the hurdle for building a server has been lowered significantly. Furthermore, Integrated Development Environment (IDE) platforms are also provided as Platform as a Service (PaaS). These systems are designed for system developers focusing on the core applications.

One example of this kind of platform is the system called Zoho Creator [14] providing an IDE for creating custom Online Database Applications. Another example is the system called COSCA [15] that provides a component-based PaaS system. A famous example is Google App Engine. App Engine is designed for developers developing applications using the program language Java, Python or Go and to run the applications on Google's infrastructure [13].

We choose the platform FOCAPLAS. This platform is different from those introduced above because it is designed for developing applications by spreadsheet formulas. By using this platform, cloud services for end users can be developed by service developers who have only spreadsheet knowledge. A platform similar to FOCAPLAS is called XCute [16]. Different from FOCAPLAS, XCute only provides Microsoft Excel ® based Web service.

In this paper, we focus on how to develop a cloud service based on FOCAPLAS, which requires mechanisms of user identification and database management. The weight control cloud service is a good developing target because all those mechanisms are required. Furthermore, it is also required to have the ability of receiving formulas posted by experts who create the formulas based on their knowledge, and receiving weight data posted by end users who use the service and feedback analysis results back to the end users based on the calculation results of the formulas.

III. THE FORMULA CALCULATION PLATFORM

The Formula Calculation Platform Service (FOCAPLAS) is a cloud service building platform that we developed. By using this service, a service provider can build a cloud service by posting formula contents like posting document contents to a blog server. If the providers want to update their services, what is required is to post new formula contents.

We are building several cloud services and testing the services for validating the effectiveness of the platform. In our testing experiments, we built a computational service, a user authentication service, a social networking service and a Web Application Program Interface (API) service. Compared with the other services similar to ours, the biggest feature of our service is that it provides a mechanism by which cloud services are built based on formulas posted by providers.

Our platform is used in the way as follows. A service provider uploads formulas to the platform server. The uploaded formulas are parsed on the server and executable program codes are generated automatically. When a user of the service connects to the server, a data submission form is sent to the user from the server side, or an application-software connected to the server is called running by the server. Data submitted by users will be calculated based on the provider-uploading formulas. Calculating results will be sent back to the users directly or indirectly. Here, indirectly sending the calculating results means that they are used as a part of query to search related data from database. In the services like the weight control service, recommendation services, etc., the retrieval results, for example, comments or recommendations will be sent back to users.

In Fig. 1, the structure of the platform is illustrated. The platform is composed of a formula parser, a formula calculator, a database management system, data storages, an input/output device and interface specification interpretation equipment.

The formula calculator receives the data sent by the service-users from the input/output device, computes received data and returns the calculation results back to the users. The formula parser parses formulas sent from service providers via the input/output device, generates executable codes and stores them into the database system.

The database management system is designed to provide functions of reading and storing six different kinds of data: (1) executable codes generated based on uploaded formulas, (2) user data, (3) user registration information, (4) response data, (5) user interface information and (6) Web API information. The interface device is designed for generating input and output forms or Web APIs defined by the serviced providers in the format of html, xml, etc. and sending them to the input/output device.

A. *The Formula Parser*

The formula parser parses formulas posted by service providers and generates the following results: service Uniform Resource Locator (URL), input/output forms, application interfaces, and executable codes.

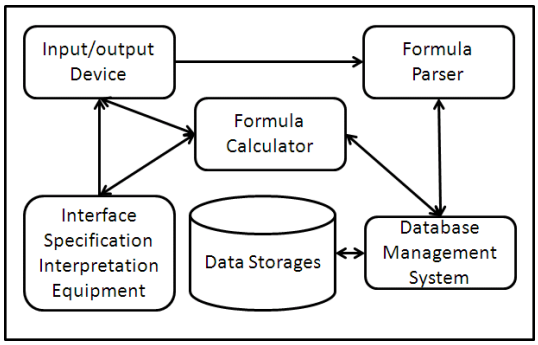


Figure 1. The structure of the formula Calculation Platform

Provider-posting formulas are opened to publics as cloud services. The formula parser can parse the formulas uploaded in spreadsheet, csv and text formats. Provider-posting formulas are parsed by the formula parser and converted into program codes that can be executed by the formula calculator.

B. The Formula Calculation Service

There are two types of formula calculation services. One is the kind of services that do not require user identification, such as "Converting Japanese calendar to the A.D. calendar", "Loan calculation", etc. The process of the service is performed in the way that user's submitted data are calculated and calculation results are feedback to the users. During the calculation process, the user identification is not required.

Another one is the kind of services, which require user identification, such as "Cumulative intake of calories", "Weight changes". The processes of this kind of services rely on historical data submitted by users. During the service processing, when new user-submitting data is received, user's previous submitted data is searched from the database system, and the current data is calculated together with the previous data.

In this paper, we discuss the latter case, in which user identification is required during the process of services. In our system, user IDs and user IDs associated data can be obtained from the database. However, we still need to develop a mechanism that connects formulas and user identification. In order to provide the healthcare service, we represent formulas for writing user-submitting data into the database and reading them back from the database based on user IDs.

C. The Database Management System

The database manage system manages a public database and user personal databases. The public database is accessed without the requirement of user identification. The user identification ID is required when a private database is accessed.

The database management system provides two kinds of management methods managing the private databases. The first kind of the method is to manage the database in which user IDs are used as the keywords during the database accessing. Another method is to manage individual user's databases. In this method, user's personal data are separately stored into different databases. When a user's personal database is accessed, an identification key is required.

We select the latter method to provide the healthcare service. A user's personal database is created when the user ID is created. If a user's personal database is not accessed during a period of times, it will be deleted. We will represent the formulas for personal database creation and deletion in section V.

IV. CREATING A WEIGHT CONTROL CLOUD SERVICE BASED ON THE FORMULA CALCULATION PLATFORM

In this section, we first describe the issue that have to be solved for creating the weight control cloud service. Then, we will represent our proposal to solve the issue after describing the related techniques.

A. The Issue for Creating the Weight Control Cloud Service

The previous user's submitted data are important when the formulas are executed during the process for providing the weight control service. However, the mechanism is required to connect user IDs, correlated to the submitted data, with the formulas uploaded from service providers. With this mechanism, the user's data can be identified under user IDs. In this way, we can find whether they are submitted by multiple users in a same day or submitted by a same user in different days. In addition, as user IDs are not filled in formulas, this mechanism helps to search user's submitted data connected to the user IDs from the database system.

B. Obtaining the User Identification Information

The data submitted to the platform server contains user identification information. Although user identification information is not contained in the data submitted to the server at the first time, it can be generated in the server side and saved in the user side. Thereby, the user identification information can be sent together with the data submitted from the next times. The user identification information can be stored in the user side in directly or indirect ways. Another method is to let a user fill a form with identifying information, such as user's login name and password. When the user identification information is used to generate a user ID, user-submitting data can be read and stored in the database under user IDs, which is used as the query keys. That is, user's previously submitted data can be obtained based on the generated user IDs.

C. Database Functions of the Platform

Different from the formula calculator, the database management system has the ability to read and write user IDs and user data, from and to the data storages. That means user data can be processed individually for each user. Utilizing this ability, it is possible to perform calculation with the previously submitted data, which can be obtained through the database management system, by sending them to the formula calculator. In the platform, formulas posted by the service provider are executed as different threads. That means user data will not be mixed during the formula execution processes.

D. Our Proposed Method

We define the keyword that appears in the database as "*k"* and its related database operation function as "*Sel*".

$$d_k \underset{Key=k}{=} Sel(Data) \quad (1)$$

In formula (1), the keyword is defined as "*k*", the field of the keyword is defined as "*Key*" and the data field is defined as "*Data*". Data stored in the database is defined as "d_k" and the operation function for reading data from the database is defined as "*Sel*".

 ISBN: 978-1-61208-280-6

An inverse function of "*Sel*", Sel^{-1} is defined for storing user data into the database.

$$Data \underset{Key=k}{=} Sel^{-1}(d_k) \quad (2)$$

During the processing of formulas (1) and (2), "*k*" is used as the keyword for reading and writing data, from and into the database. Formula (1) is used to read previously submitted data, "d_k", from the database and formula (2) is used to store new submitted data to the database.

We also defined a formula to separate user identification information "*i*" and user data "*D*" from the user-submitting data "D_i".

$$\langle D, i \rangle = Sep(D_i) \quad (3)$$

In formula (3), the data submitted by user "*i*" is defined as "D_i". During the processing of formula (3), user identification ID "*i*" and user data "*D*" are separated from "D_i".

As user identification information "*i*" can be used as the keyword for reading/writing data from/to database, it is possible to get the previous submitted data associated with the user ID. That is, data "*D*" can be read from or written to the database under user ID. Therefore, service providers only need to fill the data "*D*" in their formulas. User management is not required to be considered by the service providers.

We utilize the mechanisms of the formula calculator and the database system of the Formula Calculation Platform as the developing platform. We created the Weight Control Service to verify advantages of the platform. Based on our experiments, it is clear that cloud service with user identification can be developed without program coding. Developing time for the service is greatly reduced.

In order to protect user privacy, we selected the anonymous authentication. In our method, when a user connects our service at the first time, an anonymous user ID is created in the server side and sent to the user. The anonymous user ID is opened for mobile terminal application development. The system is also designed to accept manual inputted user ID. We also designed a mechanism to process data omissions based on the user's previously submitted data. In the service, graphs of the weight changes are generated and sent to users as it is known that showing the graphs of weight changes is helpful for weight control [17].

V. SERVICE CREATION AND EXPERIMENTS

A. Formulas for Creating the Service

We performed demonstration experiments to confirm the feasibility of the proposed method. We prepared two types of interfaces, a Web API interface for mobile devices and a HTML interface for general-purpose Web browsers. In order to generate the anonymous user ID, we create a vector, **V**, with 26 English uppercase characters.

$$\mathbf{V} = (A, B, \cdots, Z) \quad (4)$$

Each element of the vector is defined by the formula *V(i)*. In the formula, "*i*" is an integer in the range from 1 to 26. For example, *V(1)* indicates the letter "A" and *V(26)* indicates the letter "Z". An anonymous user ID is defined by a vector with "*n*" elements. Each element is a randomly selected English uppercase letter. We use the following formula to generate anonymous user ID, where rand(1…26) is a random number generator generating a number from 1 to 26.

$$ID = \sum_1^n v(rand(1 \cdots 26)) \quad (5)$$

We set "*n*" to 6 in our service. User's daily measured weight data are stored in the database system. In many cases, it is impossible letting a user input weight data every without omission. Various cases can be considered which will result to data omissions. For example, during business trips, it may be impossible to submit weight data. Or a user just does not want to submit the data because of overtime working. Considering these cases, data omissions are irregular. We use Newton polynomial [18] to interpolate the omitted weight data as it has a characteristic that the interpolated data can be reused for another interpolation. For the omitted dates, we use the linear interpolation [18] to interpolating the omitted dates with 1-day interval. Fig. 2 shows an interpolation result. In the figure, weight data are submitted on July 1, 2 and 7. The omitted dates on July 4, 5 and 6 are interpolated.

As shown in Fig. 2, in order to reduce the amount of computation on interpolation calculation, the service is designed to store data up to 7 days.

If the user identification ID, which is separated from the user-submitting data, is not stored in the database, it will be stored into the database as a new ID. We define the user identification field as *ID*, and the user ID stored in the database as ID_d. We use the following formula to determine whether the user ID is stored in the database or not. If it is not stored in the database, ID_d is defined as $ID_d = 0$.

$$ID_d \underset{Key=ID_k}{=} Sel(ID) \quad (6)$$

If the user identification ID, ID_k is not stored in the database, that is, $ID_d = 0$, we use the following formula to store it into the database.

$$ID \underset{Key=ID_k \cap ID_d=0}{=} Sel^{-1}(ID_k) \quad (7)$$

In the formula, "∩" is defined as the logical product.

We use formula (8) to store user's weight data into the database, where W_k is the user's weight data and *W* is the field of weight data.

$$W \underset{Key=k}{=} sel^{-1}(W_k) \quad (8)$$

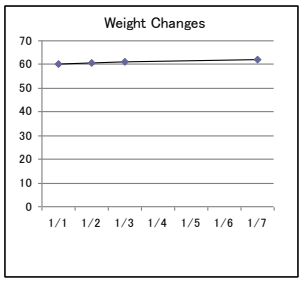


Figure 2. Interpolation result for 3 days

Formula (9) is used to read the weight data from the database.

$$W_k = \underset{Key=k}{Sel}(W) \tag{9}$$

B. Formulas for the Database Operations

Formulas for database creation and deletion are supported by the Formula Calculation Platform. In order to create a database, a field name vector, **N**, and a field character vector, **C**, are required. The database creation formula is defined as "*Cre*".

$$Cre(\mathbf{N}, \mathbf{C}, name) \tag{10}$$

In the formula, "*name*" is the name of the created database. When the vector **N** or **C** is an empty vector, the database "*name*" will not be created.

The user personal database is crated in the case that the user identification ID, ID_k is not stored in the ID database (ID_d=0). As we define all the user personal databases having the same fields, we define two common vectors, $\mathbf{N}_{cm}$, $\mathbf{C}_{cm}$, for the database creation. After the user identification ID, ID_k, is created, the user personal database will be created based on the following formulas, (11), (12) and (13).

$$\mathbf{N} = \underset{ID_d=0}{\mathbf{N}_{cm}} \tag{11}$$

$$\mathbf{C} = \underset{Id_d=0}{\mathbf{C}_{cm}} \tag{12}$$

$$Cre(\mathbf{N}, \mathbf{C}, ID_k) \tag{13}$$

Based on the formulas (11) and (12), for a user ID that is already generated and stored in the ID database, the vectors **N** and **C** are the empty vectors. Therefore, new user database will not be created.

The formula for database deletion is defined as "*Del*". A personal databases will be deleted which are not accessed more than 7 days. In the "*Del*" formula, formula (14), a database name, *name*, is required.

$$Del(name) \tag{14}$$

In the formula (14), if *name* is an empty string, the function will not be executed.

We use the following formula to control the delete operation. In the formula, the user ID log is used. When an ID is not active more 7 days, the ID will be sent to the string variable, *dn*. When the no-empty string variable is sent to the formula *Del*, the database, which name is stored in the variable *dn*, will be deleted.

$$dn = \underset{ID_k|\text{unactive log}=7}{ID_k} \tag{15}$$

The personal database is deleted by using the following formula.

$$Del(dn) \tag{16}$$

C. Checking the Operation of the Created Cloud Service

The operations of reading and writing user identification ID, and user's submitted weight data are check. The execution of the uploaded formulas for weight control is also checked.

In our experiments, we first checked whether a new ID key is created for the beginning user or not, and whether the new ID key is stored in the database or not. After that, we confirmed whether the user-submitting data is stored under the ID key into the database or not. The operation check result is shown in Fig. 3. We confirmed that a new ID key is created for the beginning user during the data submitting. For the subsequent data submitting, the user identification information is separated from the submitted data and stored into the database.

We also confirmed that new submitted weight data are correctly stored into the database under the already existed ID key. Fig. 4 is the result of the operation check.

We confirmed interpolation for the data omissions. Fig. 5 is the result of the operation check. As shown in Fig. 5, there is a data omission on January 3. The interpolation works correctly as shown in the figure. We also check the operation of the interpolation for the continuous data omissions. Fig. 6 is the result of the operation check. There are data omissions on January 3, 4 and 5. As shown in the figure, the interpolation works correctly.

We verify the operation of the service used by mobile terminals. Fig. 7 is the result of the operation check. The operation check was carried out by using Android mobile cell phones.

XOONUE		
recDate	**myWeight**	**myMemo**
2013/01/01	60	

Figure 3. The operation check result on new ID key generation

XOONUE		
recDate	**myWeight**	**myMemo**
2013/01/01	60	
2013/01/02	60.1	
2013/01/04	60.5	
2013/01/05	60.7	
2013/01/06	60.7	
2013/01/07	60.5	

Figure 4. New weight data are stored under the ID key

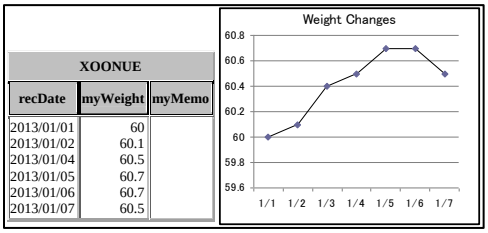


Figure 5. Interpolation for the one day omission

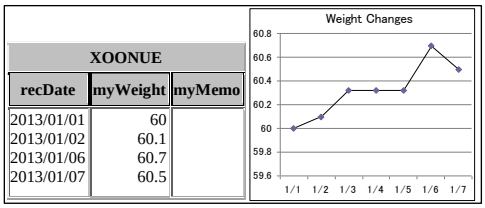


Figure 6. Interpolation for the continuous data omission

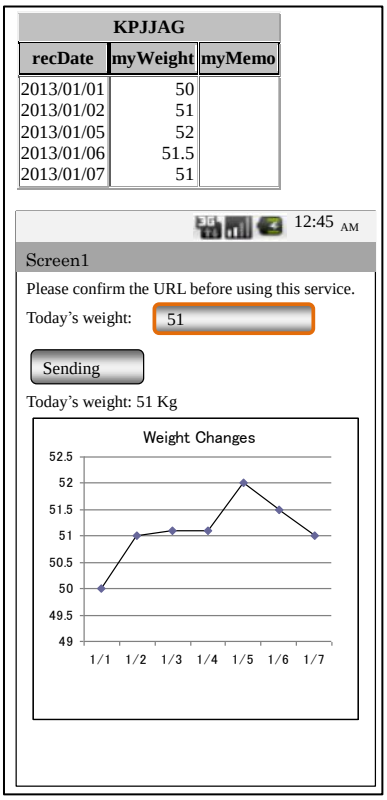


Figure 7. The operation check on an Android mobile cell phone

VI. TECHNICAL DETAILS ABOUT THE COMMUNICATION PROTOCOLS

In this section, technical details about the communication protocols are presented. In the Formula Calculation Platform (FOCAPLAS), themes are provided to the cloud service providers. Two different types of themes are provided. The first one is designed for Web browsers. The second one is designed for Web API application developers. Cloud service providers select the themes that are suitable to their services. Based on the selected themes, service providers design and upload their formulas. By using the selected themes, cloud services are performed in the following steps:

(1) Users' service requests are represented as arguments of program functions, which are generated based on the providers' functions.

(2) The program functions are called and executed according to the service requests.

(3) Return values of the program functions are formatted based on the design of the selected theme and sent back to the cloud service users.

In the FOCAPLAS, we use the Hypertext Transfer Protocol (HTTP) as the data communication protocol. We have designed three kinds of themes: (1) using GET method only, (2) using both GET and POST methods and (3) using GET, POST and PUT methods. The themes including JavaScript codes are also designed and created.

In the FOCAPLAS, an assignment operator is designed and created. Request messages are assigned to the arguments of the program functions by the assignment operator. Response messages are generated based on the return values of the program functions in the format defined in the theme, which is selected by the service provider.

To send back response messages back to the Web API applications, character strings are enclosed by tags, which indicate start and end positions of a string. According to these tags, a character string, which will be displayed in a label, can be extracted from the return message. Furthermore, an image, which will be displayed in a canvas, can also be extracted from the return message by using the tags.

The tags are user definable. We defined three kinds of tags for the weight control cloud service. The first kind of the tags is used for the string display, which is used to extract the character string that will be displayed in a label. This kind of the tags is referred to as "tags for label."

The second kind of tags is used to extract the graph URL from the return message, which shows the daily changes of weights. This kind of tags is referred to as "tags for information resource."

The third kind of tags is used to extract user identification strings from the return message, which is used for sending user ID's associated data to the server. This kind of tags is referred to as "tags for ID."

The tags used in the weight control service, with which the operation check result is shown in Fig.7, are summarized as follows:

(1) Tags for label:
Start tag: LabelStart
End tag: LabelEnd
Example:
LabelStart
Today's weight: 51Kg
LabelEnd

(2) Tags for information resource:
Start tag: ImageStart
End tag: ImageEnd
Example:
ImageStart
http://weitght.control.jp/AIEGOP.gif
ImageEnd

 ISBN: 978-1-61208-280-6

(3) Tags for ID:
Start tag: CodeStart
End tag: CodeEnd
Example:
CodeStart
KPJJAG
CodeEnd

VII. CONCLUSION AND FUTURE WORK

In this paper, we represented a method for creating cloud services based on our developed Formula Calculation Platform Service. In the method, we proposed formulas for separating user identification information, and formulas for reading and writing user-submitting data under the user identification information. We created a weight control cloud service to evaluate the efficiency of the proposed method. Our evaluation experiments were performed on three subjects.

The first subject is to evaluate the execution of the uploaded formulas. In our weight control service, we proposed individual personal-based formulas. The cooperative operation of the proposed formulas and the database manager system is a key point to confirm the effectiveness of the proposed method. We confirmed that the proposed formulas are executed correctly based on our experimental results.

The second subject is to confirm the operation of the user identification ID. We confirmed that new keys are created to the users who first used the service. We also confirmed that user identification IDs are correctly separated from the user-submitting data into the user identification ID and the user weight data. The uploaded formulas correctly worked with the server network management system. Based on the experimental results, our proposed formulas are correctly executed.

The third subject is to test the uploaded formulas working as a cloud service. It is extremely important that these formulas work together as a system providing the required cloud service. We confirmed that the weight control cloud service that we created is running correctly.

Based on the experimental results, we confirmed the effectiveness of our proposed method. That is, we have reached our goal of developing a cloud service for end users using formulas only. Our method is different from the other methods [13][15][19] because our method is suitable for the cloud service builders who have less ICT techniques. We demonstrated that cloud services that need the mechanisms of user identification and database management can also be developed by using formulas presented in this paper. By using FOCAPLAS, the overhead of developing cloud applications is greatly reduced and the hurdle for developing cloud applications is also lowered significantly.

As our future work, we will present more applications with more technical details. As a quickly developing technology, many new method and techniques are presented in the cloud computing research area. We are taking a great interest in the relative research works, especially in the area of PaaS and mobile applications working together with cloud computing.

REFERENCES

[1] NEC Corporation, “Health management cloud service”, http://www.nec.co.jp/solution/cloud/service/saas_common/healthcare.html, (in Jaspanese) (accessed 2013-02-12).

[2] NTT Resonant Inc., “A cloud-based healthcare services: ‘goo body log’ ”, http://pr.goo.ne.jp/detail/1670/, (accessed 2013-02-12).

[3] Ryobi Systems, “Health Medical Recorder (General health managing and supporting system)”, http://rcs.ryobi.co.jp/solution/by-company/rs-health-business/item/40-kenko-karte.html, (in Japanese) (accessed 2013-02-12).

[4] Chen Lab., “FOCAPLAS”, http://www.chen.ic.kanagawa-it.ac.jp/focaplas/index.html, (in Japanese), (accessed 2013-05-19).

[5] Maryam Alavi and Dorothy E. Leidner, “Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues”, MIS Quarterly, Vol. 25, No. 1, Mar., 2001, pp. 107-136.

[6] Yasushi Kiyoki, Takashi Kitagawa, and Kayoko Kurata, “An Adaptive Learning Mechanism for Semantic Associative Search in Databases and Knowledge Bases”, Information Modelling and Knowledge Bases, VIII, IOSW Press, May., 1997, pp. 345-360.

[7] Takafumi Nakanishi, Koji Zettsu, Yutaka Kidawara, and Yasushi Kiyoki, “An Interconnection Method for Heterogeneous Knowledge Bases by Utilizing Wikipedia”, 20th Technical Committee on Semantic Web and Ontology, http://sigswo.org/papers/SIG-SWO-A803/SIG-SWO-A803-04.pdf, SIG-SWO-A803-04, Jan., 2009, (in Japanese).

[8] Masakatsu Mori, Ryoichi Ueda, Nobutoshi Sagawa, and Hiroko Sukeda, “KaaS (Knowledge as a Service) for Value-added Business Creation”, Hitachi Review ,vol. 91, no. 07 600-601, Jul., 2009, pp. 56-59, (in Japanese).

[9] Minako Toba, Yasuhide Mori, and Daisuke Tashiro,”Proposal of “Knowledge Recommender” the Business-knowledge Sharing System and its Application to Building-energy-management Service”, IPSJ Transactions vol.53 no.1, Jan., 2012 , pp.149-162, (in Japanese).

[10] Freshmanmoney.com, “Income/Resident Tax Calculator, Japan”, http://en.freshmanmoney.com/tax/form/calc_form.html, (accessed 2013-02-12).

[11] Real Estate Information in Hiroshima, “You can calculate your house loan”, http://home.gr.jp/athome/file/loankeisan.html, (in Japanese), (accessed 2013-02-12).

[12] Amazon, “AWS”, http://aws.amazon.com/, (accessed 2013-02-12).

[13] Google, “Developers”, https://developers.google.com/appengine/, (accessed 2013-02-12).

[14] Zoho, “Creator”, http://www.zoho.jp/creator/, (in Japanese), (accessed 2013-02-12).

[15] Steffen Kächele, Jörg Domaschka and Franz J. Hauck, “COSCA: an easy-to-use component-based PaaS cloud system for common applications”, In Proc. of the First International Workshop on Cloud Computing Platforms, CloudCP’11, New York, USA: ACM Press., pp.1–6.

[16] Microlab Co. Ltd., ”XCute”, http://www.microlab.jp/index.htm, (in japanese),(accessed 2013-05-15).

[17] NHK, “Gatten, diet chart”, http://www9.nhk.or.jp/gatten/dietclub/, (in Japanese), (accessed 2013-02-12).

[18] Raymond W. Southworth, and Samuel L. Deleeuw, “Digital computation and Numerical Methods” McGraw-hill Book company, New York [“Mathematics for computer II - Numerical analysis-“, KYORITSU SHUPPAN CO., LTD.,pp.276-294, October, 1975, Japan] (in Japanese)

[19] M. Reza Rahimi, Nalini Venkatasubramanian, Sharad Mehrotra, and Athanasios V. Vasilakos, “MAPCloud: Mobile Applications on an Elastic and Scalable 2-Tier Cloud Architecture.”, 2012 Utility and Cloud Computing, IEEE Fifth International Conference on, Nov., 2012, pp. 83–90.

 ISBN: 978-1-61208-280-6

RESTful Correlation and Consolidation of Distributed Logging Data in Cloud Environments

Christian Pape, Sven Reissmann, Sebastian Rieger
Applied Computer Science
University of Applied Sciences
Fulda, Germany
{christian.pape, sven.reissmann, sebastian.rieger}@informatik.hs-fulda.de

***Abstract*—Due to the availability of virtualization technologies and related cloud infrastructures, the amount and also the complexity of logging data of systems and services grow steadily. Automated correlation and aggregation techniques are required to support a contemporary processing and interpretation of relevant logging data. In the past, this was achieved using highly centralized logging systems. Based on this fact, the paper introduces a prototype for an automated semantical correlation, aggregation and condensation of logging information. The prototype uses RESTful web services to store and analyze the logging data of distributed logging sources. In this context we will also present the special requirements of handling logging systems in highly dynamic infrastructures like enterprise cloud environments, which provide dynamic systems, services and applications.**

***Keywords*—Monitoring; Enterprise Cloud; Web Services; Log Analysis; Log Correlation;**

I. INTRODUCTION

The vast rise of virtualization technologies and the related wide availability of virtual machines increased the amount of logging data over the past years [1]. In addition to virtual machines themselves, cloud infrastructures, in which they are deployed, also deliver new services and applications in a fast and highly dynamic manner, producing logging data that is needed to monitor their states and service qualities. This leads to a growth of logging sources and the demand for logging systems to dynamically handle new sources and collect the corresponding data. Each new source provides detailed logging information and increases the overall amount of logging data. Typically logging data will be compressed and also anonymized at short intervals if individual-related data is included. Also, outdated log entries can be removed, but the number of logging sources (e.g., the number of virtual machines) themselves can't be reduced. For instance, in a virtualized cloud infrastructure where servers, storage and also the network is virtualized, each system, service and application should at least provide a minimal set of logging data to allow an effective analysis of the status and relevant events during service operation.

To support this analysis and evaluation across logging information originating from a large number of different distributed source systems, correlation techniques offer a way to group similar systems and applications. Furthermore, correlation can be used for the aggregation of logging data hence providing a condensation based on its relevance. In this paper, we introduce a solution to persist logging data that originated from syslog sources in a NoSQL-based database by enhancing existing solutions and using RESTful web services. For correlation and consolidation purposes, this data will also be enriched with meta information before providing the data for distributed analysis and evaluations.

The paper is layed out as follows. In the next section, the state-of-the-art of distributed logging in cloud environments is described. Section III gives examples of related work and research projects that also focus on improving the management and analysis of logging data in distributed or cloud environments. Requirements for the correlation and consolidation of logging data in enterprise clouds are defined in Section IV. Using RESTful web services and NoSQL-based storage, the prototype presented in Section V was implemented. It provides aggregation and condensation of logging data in cloud environments by correlating individual events from distributed sources. The prototypic implementation is evaluated and compared to the state-of-the-art and related work in Section VI. In the last section of this paper, a conclusion is drawn and aspects for future research are outlined.

II. STATE-OF-THE-ART

The following sections give an overview on logging in distributed environments using aggregation and consolidation techniques for standard logging mechanisms like syslog. Also, the advantages of evolving NoSQL databases, which are typically backed by RESTful webservices, are outlined.

A. Distributed Logging in Cloud Environments

Current cloud service providers offer a variety of monitoring mechanisms. For example, Amazon AWS and RackSpace both provide monitoring and alarms for their virtual machines. In the basic version, these services monitor several performance metrics (e.g., CPU, I/O and network utilization). Advanced versions (e.g., Amazon CloudWatch) allow the customers to check the current status of services running in the virtual machines and to define custom metrics and alarms that can be monitored using individual APIs of the cloud service provider. While these APIs could be used to send specific events and alarms, there is no specific service to handle the aggregation, correlation and management of logging data generated and provided by the operating systems and services running in the virtual machines. Furthermore, the individual APIs currently vary from provider to provider. Hence it is not possible to use a unified monitoring across different

 ISBN: 978-1-61208-280-6

cloud service providers. This also hinders the establishment of enterprise clouds that should allow the integration of private or hybrid cloud services operated by public cloud service customers, as these solutions again use individual monitoring techniques. An appropriate standard to address the issue of a cloud service provider independent open logging standard, is currently in the works at the IETF [2].

Until such open standards are available, distributed logging in cloud environments could be carried out by developing specific logging mechanisms for the infrastructures, platforms or applications (IaaS, PaaS, SaaS) used in the cloud. The drawback of this approach would be the effort that is needed for the software development and maintenance. Moreover, the individual APIs developed by the customers are likely to need a migration to upcoming cloud logging standards in the near future. Therefore, the more appropriate approach could be to extend existing and established logging services to support distributed cloud scenarios. The de-facto standard logging service offered in every predefined Linux-based virtual machine image by existing cloud providers is syslog, which is described in the next section. As a matter of fact, syslog is also the basis for the upcoming Internet-Draft [2] focusing on cloud-based logging services. Logging data can be stored and structured in NoSQL-based databases using RESTful web services as described in Section II-C.

B. Log Aggregation and Consolidation with Syslog

Syslog [3] defines a distributed logging solution for generating, processing and persisting host- and network-related events. Since its introduction, the syslog protocol evolved into the de-facto standard for the processing of logging events on UNIX-based systems and several network devices. A syslog message consists of multiple parts. First part is the so called PRI part, which contains a numeric priority and the facility that generated the message. Second part is a header, which includes a timestamp and the hostname of the producer of the message. The latter allows the grouping of different messages originating from the same individual machine. The closing MSG part consists of the message itself and can also include additional informations like the id of the process that produced the event. In a default configuration syslog messages of a host system are stored in files on the host's local filesystem. As outlined above, the impact of virtualization technologies and the corresponding growth of logging sources indicate that a centralized collection and analysis of syslog data is of essential importance. Otherwise, an overall rating of nearly identical messages originating from different sources would be a difficult task.

A centralized logging infrastructure and the utilization of relays to cascade logging servers in large environments, were also design goals of the syslog development. Originally, syslog [3] defines the User Datagram Protocol (UDP) to transport messages. Today the reliable Transmission Control Protocol (TCP) is preferred [4]. Also, additional security features like Transport Layer Security (TLS) assure integrity and authenticity of the data during the transport [5]. Figure 1 shows an example of a centralized logging environment.

The rsyslog server [6] provides an open source implementation of the syslog protocol and is among other solutions like

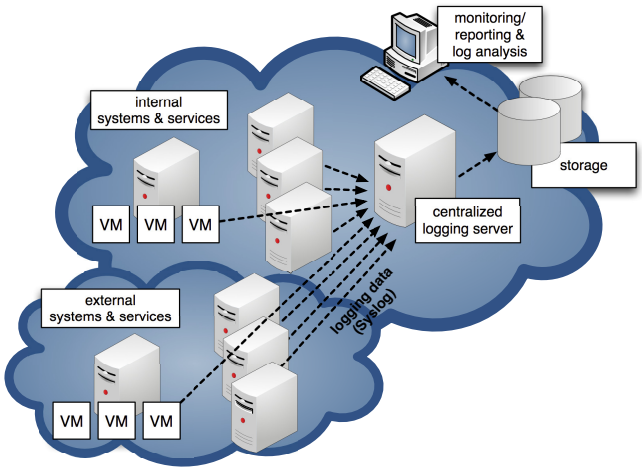


Fig. 1. Centralized logging of distributed systems and services

syslog-ng one of the most popular syslog servers. Also, a large number of plugins are available for rsyslog to support different message normalization techniques and new storage backends like MongoDB, HDFS or ElasticSearch. For these reasons, the popularity of rsyslog increased over the past years. Therefore, we use rsyslog in our research to provide centralized logging of syslog messages.

C. RESTful Log Management utilizing NoSQL Databases

The term NoSQL, standing for Not only SQL, refers to a type of databases that became an interesting alternative to SQL databases over the last couple of years. Although there are different implementations of NoSQL databases that fit different needs, they all share one aspect: They are not relational databases. A reason for the popularity of such new databases might on one hand be their performance. On the other hand the new requirements on storing unstructured data have changed with new concepts like BigData and full-text search. While relational databases demand the structure of the data to be specified when creating the database, NoSQL databases do not have such a need, allowing the structure of the data to be modified or extended at any time. Also, creating a relational database for data that does not easily map into a table-layout (e.g., different log formats from distributed sources) is not easy. Regarding the storage of logging data, the most important issues are the performance of the database system, especially as in many cases there is a tremendous amount of data to be stored, and the flexibility to add new log sources that may introduce new data structures.

When looking at the different approaches of NoSQL databases, three main types can be identified. Key-Value-Stores allow to store unstructured data simply in form of key-value pairs. These databases achieve high performance when querying for keys, but are not very well suited to perform searches on the stored values. Going one step further, column-oriented databases allow the storage of closely related data in an extendable column. Besides using a column layout, they are not bound to the restrictions of the highly structured table layout SQL uses, but also allow to store data in a more detailed structure compared to key-value pairs. A third type of NoSQL

 ISBN: 978-1-61208-280-6

databases is referred to as document-based datastores. These types of databases are able to store collections of documents, each of which having a completely independent structural layout. The structure of new documents can be extended at any time, meaning that documents may consist of any number of key-value pairs of any length. Most of the document-based databases provide a RESTful web service interface allowing to store and retrieve documents using the JSON-standard. Therefore, document-based datastores provide a high degree of flexibility and interoperability.

Regarding the requirements of storing logging data, not all of the previously mentioned NoSQL technologies are suitable for a centralized logging data storage. Key-Value stores as well as column-oriented databases allow highly efficient queries for the data using their keys, while not being suitable for doing full-text searches and correlation on the stored data. Document-based datastores in contrast, allow highly efficient search queries on the full data and also efficient queries using the documents' keys. For our work we used ElasticSearch [7] as a document-based NoSQL datastore, which also offers a high-performance, full-featured text search engine based on Apache Lucene.

The decision to use REST instead of SOAP for our web service was driven by two reasons. First and foremost, the ElasticSearch search and analytics engine primarily offers REST APIs, which mainly use JSON [7]. Also, the rsyslog daemon we chose offers corresponding output modules. Second, the usage of REST perfectly fits the logging in the cloud environments we evaluated, because it is not as strictly tied to XML as SOAP [20]. Therefore, the overhead to transfer logging messages between the logging server and the prototype to correlate and consolidate the logging data, that we present in this paper, can be minimized. Otherwise, for each syslog message being sent to our prototype, the plain text logging data would need to be embedded in an XML structure. While it would have been possible to address this issue, e.g., with SOAP Message Transmission Optimization Mechanism (MTOM), this would still increase the overhead of requests and responses from the the logging server, which also leads to a decrease in performance. On the other hand, SOAP would provide alternative transport protocols and a strict definition of the interface and used data types [21]. The latter ones are not an issue for this paper, as data types and interfaces are already defined by rsyslog output modules and ElasticSearch. Overall, the simplicity of REST [20] allows a rather light-weight implementation for the communiation needed between the central logging server and the prototype that we will present in Section V.

III. Related Work

The challenge of persisting and evaluating decentralized logging data has been in the focus of many research publications. For instance, the evaluation of decentralized logging informations in IaaS, PaaS and SaaS cloud environments were described in [8] and [9]. Also, an internet draft is in development [2], covering logging of syslog messages from distributed cloud applications. Besides the requirements by these new highly distributed applications, there is also a challenge for analysis and structuring of logging information. Existing solutions for automated log analysers only comply with some of these requirements [10]. Therefore, Jayathilake [10] recommends the structuring of logging data and to extract the contained informations. In this context, NoSQL databases are best suited for handling these variable fields. These databases provide an adaptive approach of persisting data and allow the use of different table schemata or, e.g., an document-based approach storing key-value pairs. As outlined in [11] and [12], the evaluation and rating itself can be automated by event correlation and event detection techniques. Both publications also describe the use of the correlation solution Drools, that we use for our research. Correlation techniques help to reduce (and consolidate) the logging data so that only a condensed representation including relevant information, required for analysis and evaluation, will be persisted. As described in [12], a reduction of syslog data by up to 99% is possible. A solution based on the NoSQL database MongoDB using MapReduce to correlate and aggregate logging data in distributed cloud analysis farms is described in [13]. This solution however lacks event correlation and detection techniques.

IV. Requirements for Correlation of Logging Data in Enterprise Clouds

We initially described that centralized logging environments tend to produce a tremendous amount of logging events at the central logging server. To manage the storage of all these data and provide a way to perform a fast analysis on the stored data, the use of the previously mentioned NoSQL datastores seems obvious. However, looking at the amount of data that has to be manually analysed and evaluated, the question arises whether it is possible to automate the process of evaluating the relevance of certain syslog events or even reduce the amount of data that will be stored. The latter is only reasonable if it can be guaranteed that no valuable information will be lost by the reduction of messages. In the next sections, we are going to describe our approach for automatic evaluation and reduction of syslog events in detail.

A. Correlating Distributed Logs in Enterprise Clouds

The core objective of this paper is the processing of the data provided from syslog and to identify important events of the network or individual hosts. For instance, the sequence of messages of an ongoing SSH brute force attack illustrates the demand for an automated rating of messages. During a brute force attack, the SSH daemon generates a log entry for each invalid login attempt. These messages are delivered to a centralized syslog server, indicating individual failed login attempts. However, the relative small number of events might become lost in the large total amount of syslog messages.

An IT operator analysing the logging data is not interested in displaying each individual login attempt, but rather wants to know whether the brute force attack led to a successful login. To answer this kind of question, the syslog messages must be filtered for the corresponding SSH daemons and searched for failed login attempts that are followed by a successful login. Thus, a system registering a large number of failed login attempts, and finally a successful login, might experience a successful brute force attack, while the possibility of an attack is rising along with the number of failed login attempts. The time-consuming and costly search for attack patterns like this can be simplified by an automated rating of syslog messages.

 ISBN: 978-1-61208-280-6

To identify individual messages describing similar events from different operating systems and platforms, it is required to normalize syslog data before correlating and persisting them. For the lookup of SSH login attempts, it is sufficient to examine single individual syslog messages. In order to identify a completed attack, a sequence of these matching messages must be investigated. If the conditions for a successful attack are met, it is possible to generate a new prioritized syslog message to support the immediate detection of these security threats in the network.

B. Consolidating Logging Data from Distributed Services

A second goal of our work was to reduce the amount of messages that actually get persisted into ElasticSearch. This may seem subordinate against the backdrop of increasing computing performance and concepts like BigData, but reducing the actual data still results in faster and easier analysis, even when using these new techniques. In practice, we actually see an advantage of the reduction of stored messages in long-term storage and data analysis. Such reduction techniques basically delete messages of a certain age or don't persist messages below a certain severity. However, these simple mechanisms result in a loss of valuable information, and for this reason are not practicable in our view.

Our approach first provides a grouping of messages. For example, same or recurring events are summarized. Based on those groups we are able to generate new summarized syslog messages containing a dense representation of all the valuable information of the initial messages and hence allowing us to actually drop those without loosing information. Using our solution, it is possible to reduce the amount of messages that needs to be stored at the central database server, and therefore improving the performance of the system without loosing information. Furthermore, it is possible to manipulate the severity of the newly generated messages, to even increase their value for later analysis.

An example of such a modification of syslog messages could be used to detect the previously mentioned brute force attack, that results in a flood of messages with a low priority. By generating a single message with a high priority - telling an administrator what the actual attack looked like, judging from the number of login attempts, the duration of the attack and the actual result - we produce information that helps to estimate the situation and the next steps to be taken. Also, regardless of waiving all the failed login attempts at the central database server, it is still possible to perform an exact analysis of the attack by looking into the logfiles of the actual server that was under attack.

A second example of consolidating messages would be the correlation of application access logs. For instance, in a cloud environment new machines will be spawned on demand, so several machines provide a single service in a cooperative way. An example could be a number of dynamically started HTTP servers receiving requests via a load balancer. The requests on the individual servers will be logged to the centralized syslog server, but these individual events must be aggregated, e.g., to support the decision process of starting new or stopping VMs running an HTTP server. The access log messages can be correlated to an access count per timeslot and it is also possible to count active HTTP servers by differentiating distinct logging sources. As already illustrated in the previous example it is again not necessary to persist the original access messages. The correlated logging information is useful to evaluate the load on all servers and can also be used to determine whether running machines have to be stopped or new ones need to be started.

Taking the logging information into account a thorough decision can be made that goes beyond the possiblities of network-based load balancing and failover techniques. A more generic approach would be to use Drools to count messages matching a set of rules for specific timeslots and to generate histograms for these kind of messages. This approach allows to compare different timeslots and answer questions like "Were the same number of crond jobs executed on monday and tuesday?". Also, a visual representation of these results, e.g., as presented in [14], could be possible with the benefit of easily identifying anomalies at first sight.

V. Implementation of Log Correlation and Consolidation in Cloud Environments

To facilitate the analysis and storage of logging data in distributed cloud environments, this paper presents a log correlation and consolidation prototype. The prototypic implementation uses rsyslog [6] as a central syslog server that receives and normalizes syslog messages originating from distributed sources, e.g., VMs in the cloud. After normalizing the data and hence allowing to process the data from different sources in a unified way, the logging information is serialized to JSON and sent using a RESTful web service to our prototype, which we implemented in Java. The prototype embodies a correlation engine, which analyses the messages and afterwards persists them, again using JSON via a RESTful web service, in an ElasticSearch cluster. Our implementation of the correlation is based on the Complex Event Processing (CEP) Engine Drools Fusion [15]. This engine allows the definition of rules using temporal reasoning that we use for the correlation of messages.

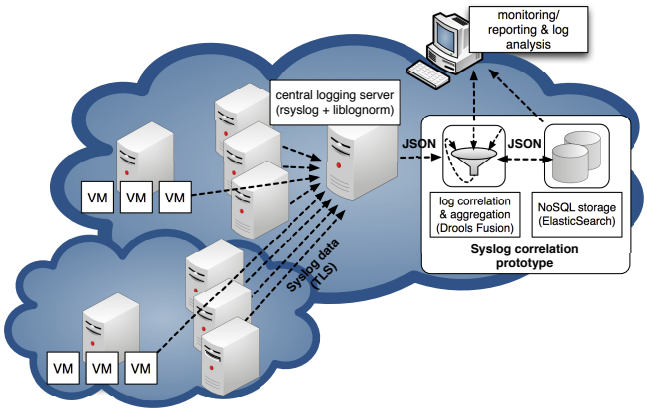


Fig. 2. Correlation and aggregation of centralized logging data

Figure 2 shows the setup of our testbed. Syslog messages are sent from the distributed clients to the rsyslog server using a secure TLS connection. The server normalizes the messages using liblognorm [16] and transmits the normalized messages in JSON format using RESTful web services

to the correlation prototype. By using REST, our prototype implements a flexible interface that can easily be used from a variety of cloud services. Moreover, the majority of the compute cloud providers offer syslog-based logging in their VMs and therefore our entire approach using rsyslog can be used to correlate the logging data across multiple and heterogenous cloud environments. Having received the messages, our prototype temporarily stores the logging data locally, but also instantly transfers them to the ElasticSearch cluster, again using RESTful JSON-based web service requests. The temporal storage is currently needed for the evaluation of the defined rules in the Drools engine, as it is configured to be done in-memory to achieve a better performance. As described in the last section of this paper, we're also evaluating persisting the messages directly and execute the correlation in the NoSQL storage to overcome the size limitation of our in-memory approach. The rules can be easily extended to provide multiple correlation and consolidation techniques. An example is presented in Section VI. Messages that did not match any of the rules will be removed from the in-memory cache. On the other hand, messages matching at least one rule are correlated, and the result is stored with a flag representing the successful correlation and a reference to original messages that have been correlated in the ElasticSearch cluster. By cyclic searching for successful correlation flags in the ElasticSearch cluster and pruning the original messages they refer to, a consolidation is achieved.

VI. Scalable RESTful Log Analysis in Cloud Environments

In this section we give an example of the usage of our solution for the correlation of logging data being generated during an SSH brute force attack as described in Section IV-A. The aim is to generate a new syslog message with a higher priority in the case of a successful SSH login immediately after a certain number of failed logins during possible SSH brute force attacks. To detect this scenario we defined the rules shown in the following listings in the correlation engine of our prototype. Hence, the detection of the successful SSH brute force attack will be done automatically by our prototype. For the correlation of the syslog messages that indicate a successful brute force attack on the password during an SSH login, the corresponding syslog messages need to be isolated and filtered from the stream of logging data originating from the syslog server. Examples for the failed and successful SSH login messages are shown in listing 1. The messages shown here have already been normalized by rsyslog and are therefore independent of the individual host that generated them.

```
% Failed password for root from 192.168.1.1
    port 34201 ssh2
% Accepted password for root from 192.168.1.1
    port 34201 ssh2
```

Listing 1. Syslog message of failed and successful SSH logins

To correlate these events, our prototype implements the rules using Drools Fusion [15], which detect messages matching a successful SSH login after a certain amount of previously failed logins (based on the message format shown in listing 1). Listing 2 contains the rules we defined for our example. The rule matching failed messages requires a success message within 1 minutes after at least 10 failed messages.

```
success : Message(
  message matches
    "Accepted password for [^\\s]* from [^\\
        s]* port [^\\s]* ssh2")
failed : ArrayList( size >= 10 ) from
    collect(
  Message(this before[0,1m] success,
    message matches
      "Failed password for [^\\s]* from [^\\
          s]* port [^\\s]* ssh2") )
```

Listing 2. Drools fusion rules to detect successful SSH brute force attacks

If the failed rule that we defined in listing 2 matches, our current implementation generates a new syslog message with the facility "security" and priority "emergency" containing the message "Possible successful SSH brute force attack". It would also be possible to postpone the persistence of the logging data until the correlation is finished, to store all message related to the attack with a higher priority, e.g., "emergency", in the ElasticSearch cluster. Another possibility would be to only persist the new message that indicates the possible SSH brute force attack with a high priority and drop the other messages that have been correlated. Messages needed for the correlation are kept automatically in the in-memory cache by Drools Fusion according to the rules we defined. Drools automatically removes messages from the cache that do not match any of the rules anymore.

VII. Conclusion and Future Work

In the previous sections, we presented a solution to automatically correlate and consolidate syslog messages containing logging data from distributed sources in cloud environments. Besides evaluating the requirements for such implementations and defining an appropriate concept, a prototype based on RESTful web services and NoSQL database storage was developed. The prototype addresses the requirements for correlation and consolidation of distributed logging sources in today's enterprise cloud environments. It supports the proper condensation of log messages by grouping individual messages. The achieved reduction improves the performance of processing and analysing logging data especially in distributed environments with a lot of systems (typically virtual machines) sending similar logging information.

Existing monitoring solutions could be enhanced to use the presented prototype as a filter improving the quality and relevance of the logging data (e.g., by using escalation techniques, traps, or sending messages regarding detected events) as shown in the example of an SSH brute force attack in Section IV-A and VI. The integration of the prototype with existing network monitoring tools (e.g., OpenNMS, splunk) is one of the next steps for our research. An interesting starting point could be their interfaces to correlate events, i.e., to perform a root-cause-analysis, that could be extended to consume relevant events that were filtered from the distributed logging data by our prototype. Another option could be to use these interfaces bidirectionally to enrich the logging information, e.g., combining information in the logging data with the location or other details from asset, configuration, system or service management. For example, expected downtimes

 ISBN: 978-1-61208-280-6

could be resolved to ignore corresponding log events in the prototype.

While this paper focuses on the usage of the de-facto network-based logging standard syslog, the prototype presented in this paper could also handle different text-based logging sources (e.g., application specific log files, log4j, etc.). A current limitation regarding the amount of logging data that can be correlated is the available memory. Theoretically, the prototype could also use data that is already stored in the NoSQL storage for the correlation to overcome this limitation. While this approach has a negative impact on performance, it could on the other hand dramatically increase the accuracy of complex correlation over long-term data. The enhancement could be easily implemented using the RESTful search API not only for the analysis but also while filtering and before persisting the logged data in the NoSQL database. In the next version of our prototype, we will implement this extension and evaluate the performance impact (regarding latency to store a log entry and overall throughput of the correlation engine). Also, balancing the load of complex correlations across multiple instances of the prototype, e.g., elastically in the cloud, could be an option. Using OpenStack we currently set up an enterprise cloud environment to serve as a scalable platform for our prototype.

Our predefined rule set outlined in this paper can easily be generalized to fit the requirements of other use cases. In our ongoing evaluation we will therefore contrast the results of our prototype to comparative work being presented in [11], [12] and [13]. Another possible topic for future research could be the integration of existing knowledge-based systems and automated reasoning as developed, e.g., for network anomaly and intrusion detection systems (IDS). Even more interesting could be the integration of existing work that has been published regarding the detection of anomalies in syslog messages. Makanju et. al. [17] are describing a promising solution to detect anomalies in logging data of high performance clusters (HPC). Administrators can confirm the detected anomalies to correlate them with error conditions and trigger a consolidation. These techniques could also facilitate the definition of correlation rules as patterns are detected without prior configuration. Syslog-based event forecasting, as described, e.g. in [18], could be another promising option for our prototype. The prototype could be used to enhance the information being evaluated to generate the forecast, but can also consume the forcasting data. This way, existing rulesets could be augmented. Furthermore, the definition of rules could be simplified by automatically deriving rules from the forecasts, which have been submitted to our prototype. To detect failures and error conditions in cloud environments this has already been proposed in [19]. We will evaluate to extend this approach to allow for the correlation and aggregation of logging data in enterprise cloud environments.

REFERENCES

[1] C. Canali and R. Lancellotti, “Automated clustering of vms for scalable cloud monitoring and management,” in Software, Telecommunications and Computer Networks (SoftCOM), 20th International Conference on, 2012, pp. 1-5.

[2] G. Golovinsky, D. Birk, and S. Johnston, “Syslog extension for cloud using syslog structured data - draft-golovinsky-cloud-services-log-format-03,” Internet-Draft, IETF, 2012.

[3] C. Lonvick, “RFC 3164: The BSD syslog protocol,” Request for Comments, IETF, 2001.

[4] R. Gerhards, “RFC 5424: The syslog protocol’,” Request for Comments, IETF, 2009.

[5] K. E. Nawyn, “A security analysis of system event logging with syslog,” SANS Institute, no. As part of the Information Security Reading Room, 2003.

[6] R. Gerhards, “The enhanced syslogd for linux and unix rsyslog,” http://www.rsyslog.com, [retrieved: 4, 2013].

[7] Elasticsearch Global BV, “The enhanced syslogd for linux and unix rsyslog,” http://www.elasticsearch.org/, [retrieved: 4, 2013].

[8] R. Marty, “Cloud application logging for forensics,” in Proc. 2011 ACM Symposium on Applied Computing, ACM, 2011, pp. 178-184.

[9] A. Rabkin and R. Katz, “Chukwa: A system for reliable large-scale log collection,” in Proc. 24th international conference on Large installation system administration, USENIX Association, 2010, pp. 1-15.

[10] D. Jayathilake, “Towards structured log analysis,” in Computer Science and Software Engineering (JCSSE), International Joint Conference on, 2012, pp. 259-264.

[11] A. Müller, C. Göldi, B. Tellenbach, B. Plattner, and S. Lampart, “Event correlation engine,” Department of Information Technology and Electrical Engineering - Master’s Thesis, Eidgenössische Technische Hochschule Zürich, 2009.

[12] M. Grimaila, J. Myers, R. Mills, and G. Peterson, “Design and analysis of a dynamically configured log-based distributed security event detection methodology,” The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, vol. 9, no. 3, 2012, pp. 1-23.

[13] J. Wei, Y. Zhao, K. Jiang, R. Xie, and Y. Jin, “Analysis farm: A cloud-based scalable aggregation and query platform for network log analysis,” in Cloud and Service Computing (CSC), International Conference on, 2011, pp. 354-359.

[14] K. Fukuda, “On the use of weighted syslog time series for anomaly detection,” in Integrated Network Management (IM), IFIP/IEEE International Symposium on, 2011, pp. 393-398.

[15] J. Community, “Drools - jboss community,” http://www.jboss.org/drools/, [retrieved: 4, 2013].

[16] R. Gerhards, “A syslog normalization library,” http://www.liblognorm.com, [retrieved: 4, 2013].

[17] A. Makanju, A. Nur Zincir-Heywood and E. E. Milios, “Interactive Learning of Alert Signatures in High Performance Cluster System Logs,” in Network Operations and Management Symposium (NOMS), IEEE, 2012, pp. 52-60.

[18] A. Clemm and M. Hartwig, “NETradamus: A forecasting system for system event messages,” in Network Operations and Management Symposium (NOMS), IEEE, 2010, pp. 623-630.

[19] Y. Watanabe, H. Otsuka, M. Sonoda, S. Kikuchi and Y. Matsumoto, “Online failure prediction in cloud datacenters by real- time message pattern learning,” in Cloud Computing Technology and Science (CloudCom), 4th International Conference on, IEEE, 2012, pp. 504-511.

[20] C. Pautasso, O. Zimmermann and F. Leymann, “Restful web services vs. ’big’ web services: making the right architectural decision,” in Proc. of the 17th international conference on World Wide Web, ACM, 2008, pp. 805-814.

[21] M. Zur Muehlen, J. V. Nickerson, K. D. Swenson, “Developing web services choreography standards - the case of REST vs. SOAP,” Decision Support Systems, 40.1, 2005, pp. 9-29.

 ISBN: 978-1-61208-280-6

Synergic Data Extraction and Crawling for Large Web Sites

Celine Badr, Paolo Merialdo, Valter Crescenzi
Dipartimento di Ingegneria
Università Roma Tre
Rome - Italy
{*badr, merialdo, crescenz*}*@dia.uniroma3.it*

***Abstract*—Data collected from data-intensive web sites is widely used today in various applications and online services. We present a new methodology for a synergic specification of crawling and wrapping tasks on large data-intensive web sites, allowing the execution of wrappers while the crawler is collecting pages at the different levels of the derived web site structure. It is supported by a working system devoted to non-expert users, built over a semi-automatic inference engine. By tracking and learning from the browsing activity of the non-expert user, the system derives a model that describes the topological structures of the site's navigational paths as well as the inner structures of the HTML pages. This model allows the system to generate and execute crawling and wrapping definitions in an interleaved process. To collect a representative sample set that feeds the inference engine, we propose in this context a solution to an often neglected problem, called the Sampling Problem. An extensive experimental evaluation shows that our system and the underlying methodology can successfully operate on most of the structured sites available on the Web.**

***Keywords*-data extraction; crawler; web wrapper; sampling;**

I. INTRODUCTION

Large data-intensive web sites contain information of interest to search engines, web applications, and various online service providers. These sites often present structural regularities, embedding content into predefined HTML templates using scripts. Regularities are also apparent at the topological level, with similar navigational paths connecting the pages obeying to a common template. In this paper, we extend the work introduced in [1], to address two related issues for capturing useful information from structured large web sites: first, the pages containing the information of interest need to be downloaded; second, the structured content needs to be extracted by web wrappers, i.e., software modules that collect page content and reorganize it in a format more suitable for automatic processing than HTML.

In general, crawlers and wrappers generation have been studied separately in the literature. Numerous tools exist for generating wrappers, with different levels of automation. They usually aim at extracting information from semi-structured data or text, and they use to that effect scripts or rule-based wrappers that rely on the structure or format of the source HTML. In some cases, wrappers are based on ontlogies or NLP techniques. Concerning the level of automation, hand-coded wrappers require a human expert, which becomes a cumbersome task for data extraction on large data-intensive web sites. Fully automatic wrappers have also been implemented [1], [2], but they necessitate considerable data post-processing and may suffer from lower accuracy. In semi-automatic wrapping, the main focus has been on learning approaches that take several positive and negative labeled examples as input.

Various tools also exist to crawl web pages and entire web sites. Popular techniques start with seed URLs and either search on the pages for hyperlinks matching certain patterns, or consider all encountered hyperlinks and then apply selection and revisit techniques for downloading target pages based on content or link structure [3], [4]. For deep Web crawling, some work has been proposed to obtain and index URLs resulting from different form submissions, e.g., [5]. However, the production of crawlers for structured web sites remains a subject with large room for improvement.

When it comes to large web data-intensive sites, it is commonly useful to extract data from a subset of the pages, in general pertaining to vertical domains. For example, in a finance web site, one may be interested in extracting company information such as shares, earnings... In this case, there is no need to download all the site's pages about market news, industry statistics, currencies, etc. Thus we propose a new approach in which the two problems of crawling and wrapping are tackled concurrently, where the user indicates the attributes of interest while making one browsing pass through the site hierarchy. The specifications are created in the same contextual inference run. Moreover, the execution of the wrappers takes place while the crawler is collecting pages at the different levels of the derived web site structure. The mutual benefits are manifested as the produced simple wrappers extract the specific links followed by the crawling programs, and the crawlers are in turn used to obtain sample pages targetted for inferring other wrappers. We have developed a working system that offers these capabilities to non-expert users. It is based on an active-learning inference engine that takes a single initial positive example and a restricted training set of web pages. We also define in this context the *Sampling Problem*, a problem often undervalued in the literature, and we show how it is mitigated by our approach when collecting a representative training set for

 ISBN: 978-1-61208-280-6

the inference process.

Throughout the paper, we consider an example web site that offers financial information on companies. We are interested in extracting from company pages data about stock quotes that is accessible in two different navigation sequences in the site topology: first, the home page displays companies' initials as links. Clicking on a letter leads to a listing of all the companies with this given initial. Each company name in turn links to a page containing the data to extract. Second, by filling and submitting a form on the home page, one reaches index pages grouping the companies by financial sector. Index pages are paginated, and by following the links provided in each paginated result list, the target company pages can be finally reached.

The rest of the paper is organized as follows: Section II presents our web site abstract model, on which we build the interleaved crawling and wrapping definitions; Section III lists and explains the crawling algorithm; Section IV defines the extraction rule classes used for extracting both data and links leading to the pages where these data are located; Section V presents the wrapper and assertion constructs that enhance our synergic crawling and wrapping tasks; in Section VI, we define the sampling problem and present our approach to collect a sample set; Section VII summarizes the results of the extensive experimental activity we conducted; Section VIII discusses related work; and finally, Section IX concludes the paper and presents some possible future developments.

II. An Abstract Model for Describing Large Web Sites

To access data from data-intensive web sites, we aim to reach these data on pages collected by a crawler using wrappers tailored to the user's information needs. We note that large web sites are composed of hundreds of pages that can generally be grouped in few categories, such that pages of the same category share a common HTML template and differ in contents. These sites also exhibit topological regularities in the navigational paths that link the pages and page categories. By capturing these regularities, we describe large web sites with an abstract model of three interrelated levels: the *intensional, extensional,* and *constructional* levels.

A. Intensional level description

Our intensional model defines two main constructs for building schemes: the *Page-Class* construct describes a set of similar pages of the same category, while the *Request-Class* construct models a set of homogeneous *requests* (GET or POST) to navigate from the pages of one Page-Class to those of a consecutive Page-Class. In our model, *Request-Classes are typed*, in the sense that each Request-Class specifies the Page-Class that it leads to.

The above concepts are illustrated in the graph of Figure 1: for our example site, the home page can be modeled by a singleton Page-Class HOME from which depart two Request-Classes, BYINITIAL and BYSECTOR, leading to two Page-Classes, INDEXBYINITIAL and SECTOR. INDEXBYINITIAL models the index pages grouping companies by initials while SECTOR describes index pages grouping companies by financial sector.

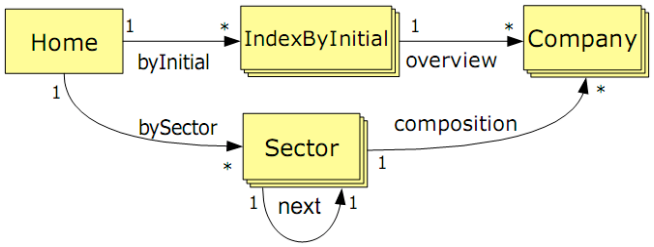


Figure 1. Intensional model example.

Both Page-Classes, by means of Request-Classes OVERVIEW and COMPOSITION respectively, lead to the last Page-Class COMPANY whose pages contain detailed information about companies, one company on each page. In particular, Request-Class NEXT models the link leading to another page instance of the same Page-Class SECTOR.

B. Extensional level description

An extensional listing of a Page-Class is a set of pages called Page-Class *support* that: (i) obey to a common HTML template and hence have similar structure; (ii) are about a common topic (each page represents a different instance of the same conceptual entity); (iii) are reachable by similar navigational paths, that is, by crossing pages that also correspond to predefined Page-Classes. Similarly, the extensional definition of a Request-Class is a set of requests called Request-Class *support* that: (i) lead to page instances of the same Page-Class; (ii) can be generated by clicking on comparable links or by different submissions of the same web form.

The two intensional constructs can be used together to build complex schemes, and to each scheme an extensional counterpart can be associated. That is, a scheme instance can be obtained by arranging the supports of every Page-Class and Request-Class involved into a graph that reflects the logical organization of the modeled site, where nodes represent page instances of a Page-Class in the scheme, while edges represent request instances of a Request-Class. We call an instance *empty* whenever every Page-Class (and hence every Request-Class) has empty support. An extensional graph for our example is partially shown in Figure 2.

C. Constructional level description

The constructional level in our model bridges the intensional and extensional descriptions by providing all the operative details needed to build the schema instances. Constructional elements are in fact the information that the

 ISBN: 978-1-61208-280-6

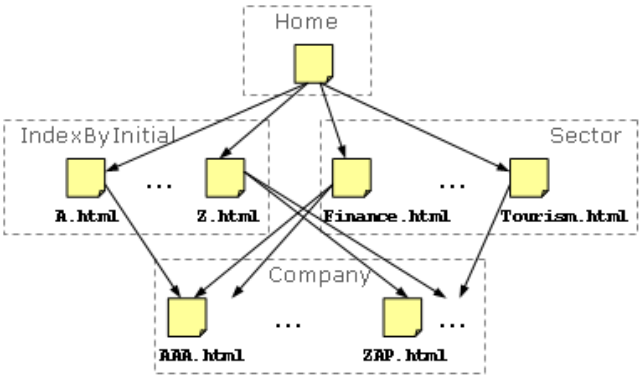


Figure 2. Extensional model example.

system needs to start from the entry page, determine the possible navigational paths to follow, proceed to subsequent pages, and extract the attributes of interest to the user. These elements consist of: (i) for the *entry* Page-Class, the set $E = \{e_1, \ldots, e_n\}$ of addresses of the pages in its support (typically the URL e of the page is sufficient); (ii) for each Request-Class r, a function er^r that builds its support, given a page instance of the Page-Class from which it departs. We call this function an extraction rule since it extracts requests from pages, e.g., the initials hyperlinks on the home page. Two other constructional elements, *assertion* and *wrapper*, are added to our model in order to enhance the Page-Class constructs. They are detailed in Section V.

III. Crawling Algorithm

Having defined our abstract model and how to build its instances, we explain how the crawler operates in this context. We first formalize the definition of navigational path, a main component for our crawling algorithm. On the intensional level, a navigational path $\mathcal{N} = (P_0 \cdot r_0^1 \cdot P_1 \cdot r_1^2 \cdot \ldots \cdot r_{h-1}^h \cdot P_h)$ is a sequence of Page-Classes and Request-Classes such that each Request-Class r_k^{k+1} is outgoing from Page-Class P_k and ingoing to Page-Class P_{k+1}, $k = 1..h-1$. It essentially corresponds to a path in the graph associated with the scheme. For our purposes, the interesting navigational paths start from an *entry* Page-Class (HOME, in our example) and reach a *target* Page-Class containing the relevant data (COMPANY). On the extensional level, the navigational path consists of *navigational trees*, where each tree is rooted at one entry page of the site and its descendants represent the pages visited by following the various requests in the support of the Request-Classes traversed.

A generalization of the user's sample browsing activity derives the navigational path definition as a sequence of constructional Page-Class and Request-Class elements. The crawler then finds all the navigational tree instances of this given path in the web site, in order to eventually download -strictly- all the pertaining pages.

Algorithm 1: Crawling algorithm based on the abstract model for large web sites

Input : A navigational path N = $(P_0 \cdot r_0^1 \cdot P_1 \cdot r_1^2 \cdot \ldots \cdot r_{h-1}^t \cdot P_h)$;
addresses $E = \{e_1, \ldots, e_n\}$ of P_0's pages;
a set of extraction rules $\{er^{r_k^{k+1}} | k = 0, \ldots, h-1\}$

Output: The navigational trees T instances of N.

Let $T = \{t_1, \ldots, t_n\}$ **be** an empty instance of N;
foreach $e_i \in E$ **do**
  add the page with address e_i as root of t_i;
for $k = 0$ **to** $h-1$ **do**
  foreach *page* $p \in \mathtt{support}^{t_i}(P_k)$ **do**
    foreach *request* $r \in er^{r_k^{k+1}}(p)$ **do**
      Let d **be** the page obtained by means of r;
      insert d as a child of p in t_i;
      insert d in $\mathtt{support}^{t_i}(P_{k+1})$;
return T;

$support^{t_i}(P)$	support of P in the navigation tree t_i
$er^{r_k^{k+1}}(p)$	set of requests associated with the r_k^{k+1} and extracted by applying the extraction rule er on page p

To do so, a crawler operates according to Algorithm 1. It starts with an input navigational path and an empty set of the corresponding navigational tree instances. Then it incrementally builds each instance by first adding a root page from the support of the entry Page-Class P_0, using to that effect the input addresses E. Subsequently, for each page p in the Page-Class under consideration, the algorithm uses $er^r(p)$ to build the support of the outgoing Request-Class r, that is, to extract on p the actual set of requests corresponding to r. These requests are sent to the web server in order to obtain new pages. The latter are added to the support of the Page-Class that constitutes the destination of the processed Request-Class. The crawler continues by iteratively picking each page p from the support of an already populated Page-Class and following its corresponding requests, once extracted. It thus incrementally builds other instances and the algorithm stops when all the requests have been sent.

IV. Specification of Extraction Rules

To perform wrapping and crawling tasks in an interrelated mode, our system infers extraction rules. These rules are used in crawling to locate the requests to be followed, and in wrapping to extract the relevant data from the downloaded target pages.

In Section II, we have already discussed extraction rules for requests. We revisit the previous definition to also

 ISBN: 978-1-61208-280-6

model wrappers for extracting relevant data from a page. An extraction rule er is then more generally defined as a function that locates and returns a set of strings s_i from the HTML code of page p: $er(p) = \{s_1, \ldots, s_k\}$.

In order to infer extraction rules, our algorithm takes as input a few positive examples provided by the user, highlighting the strings to extract (links or attributes). Eventually, only the rules compatible with collected user feedback are kept. Defining the class of inference rules to be generated constitutes then an important design choice. On one hand, a large and expressive class is more likely to include all the useful rules but may require many initial samples and a lot of user interaction before the correct rules are discerned. On the other hand, a limited and concise class of extraction rules requires less samples, and therefore less user interaction, but is also less expected to include correct extraction rules that are valid on all the pages. We address this tradeoff by opting for extraction rules based on XPath expressions and obtained from the union of three simple classes:

ABSOLUTE containing all absolute XPath expressions with positional predicates in each location step, generalized when a superset of several samples is needed

URL-REGEX obtained by filtering the rules in the ABSOLUTE class with simple regular expressions, and used mainly for extracting links where the URL value obeys to a general pattern

LABELED consisting of relative XPath expressions that make use of a fixed node considered part of the template to locate the string to extract

These classes proved to work on more than one hundred real web sites with very good performance results, while maintaining simplicity and requiring very limited user effort to discern a correct rule.

V. WRAPPER AND ASSERTION CONSTRUCTS

As mentioned, wrappers and assertions are constructs used in our constructional model to enhance the system's definitions of crawling and data extraction activities.

Wrappers are tools for extracting data values on web pages. Data wrappers generated by our system consist of rules taken from the LABELED class as they are less sensitive to variations on the page. They require that the node chosen as a reference be present in most of the target pages while being as close as possible to the data node. We associate the wrapper element with the Page-Class construct so that when crawling to build a Page-Class support, if a wrapper definition is encountered, it is executed instantly on the downloaded page.

Assertions are Boolean predicates over web pages. They are formulated as a set of XPath expressions locating valid template nodes, and a page is said to satisfy an assertion if and only if it agrees with all the XPath expressions associated with the template of its respective Page-Class. As we relax extraction rules and allow them to extract a superset of the correct links, the system can detect and discard any *false positive* pages crawled by checking whether their template matches or not all the established assertions. Consequently, assertions replace the need for adding more expressive classes of extraction rules when the crawler's performance is not satisfactory, which means fewer rules produced in response to the examples provided by the user, and fewer subsequent examples required from the user to discern the correct rules [6].

VI. SAMPLE SET SELECTION

The input required for the semi-automatic generation of the crawling programs and annexed wrappers is provided through interaction of a non-expert user with our system through a browser-based interface. For the model generation, the system tracks user navigation and derives an *internal schema* from the user's browsing activity as per the model listed in Section II, with the Page-Classes and Request-Classes of the web site along with the needed extraction rules. As for wrappers inference, the system generates extraction rules for the data selected by the user and evaluates them with the active learning module of the underlying inference engine. Then for any page with uncertainty, be it for the template or for the extracted values, the user is prompted to confirm the results, correct them, or even discard the page.

A. Sampling Problem

In order to produce correct and effective extraction rules, automatic and semi-automatic inference methods require samples with a certain quality of representativeness [6] that inexpert users cannot provide. Therefore the sample pages chosen by the tool to collect user feedback need to be representative of their corresponding Page-Classes. In addition, the navigational paths linking them together need to cover a rather wide variety of the possible paths in the selected informative domain. As a result, selecting sample pages introduces what we define as the *Sampling Problem*. This problem, often neglected, consists in determining which sample pages to choose in order to have "good" positive examples and guide an inexpert user in the inference process. Consider the following two scenarios:

Example 1: the downloaded sample pages all belong to one specific subset; say company pages from the Agriculture sector in our running example. Such pages would all contain the token "Sector: Agriculture".

Example 2: the downloaded sample pages share a particular variation in the template, such as the pages of top-ranked companies in our example containing a table with statistical data. The headers of this table constitute tokens that are not shared by other pages that do not rank first in their respective sector.

In these examples, sample pages may not be representative of all their Page-Class instances, as they contain some

 ISBN: 978-1-61208-280-6

specific tokens that are not present in the remaining target pages. This can in turn affect negatively the crawler, were wrong template assertions derived from these tokens, as well as the data extraction process, were any relative rules to be built around these tokens.

B. Resolution

The manifestations of the sampling problem are: (i) in characterizing the valid pages template, (ii) in collecting good navigational path instances, (iii) and in generating accurate extraction rules for the data of interest. In the first case, the problem is to generate a template characterizing only relevant web pages and discard "false positives". A template consists of a set of valid tokens that are present in most of the sample pages at the same XPath location. The second case relates to the need of covering the domain's navigational paths with minimum bias, and is addressed by tracking and generalizing the user's navigation at each step. The third case can then be resolved having collected a diversified page sample and derived a valid template. We propose the following approach to characterize valid page templates based on a statistical analysis of page contents and a learning algorithm requiring limited user effort:

- at each navigation level, consider all the pages obtained by generalizing selected links or form requests
- from this set of pages, which can be very large, choose a fixed percentage of random sample pages to download
- analyze tokens occurrence and data extraction results on the sample pages to train the classifier
- apply uncertainty sampling techniques to select query pages to propose to the user for feedback
- update tokens set, assertions, and extraction rules from user feedback

With this technique, our system is able to collect and use a representative subset from the site pages to infer performing wrappers and crawlers relevant to the user's needs. At execution time, the constructional details, as described in Section II-C, recorded in XML format, are used to guide the interleaved crawling and wrapping on the large web site.

VII. Experiments

In this section, we summarize our experiments conducted with the system prototype, called CoDEC, implemented as a Mozilla Firefox extension. We used our prototype for generating specifications and executing them on real web sites to analyze the performance on a wide variety of heterogeneous topics, page templates, and navigational paths. We evaluate our experiments by computing the F-measure as

$$F = 2\frac{P * R}{P + R} \quad (1)$$

This value ranging between 0 and 1 reports the harmonic mean of the precision *P* and recall *R*. For crawling, it evaluates the set of downloaded pages as compared to the set of actual target pages. Whereas for wrapping, the F-measure evaluates the values extracted by the generated rules with respect to the correct set of attribute values on the target pages. Larger F-measure values imply better results.

Table I sums up the experiment results of our crawling techniques on 100 different sites belonging to various domains. The simplest class ABSOLUTE was sufficient to extract most of the links leading to target pages. In few cases, where other links leading to non-target pages were located very close to the correct links, URL-REGEX extraction rules improved the precision by discarding the links that do not match an inferred URL pattern. One crawling limitation was the inability to discard target pages with the same template but different semantic entities, such as pages for coaches downloaded with those of players in a soccer web site.

Table II summarizes the results of testing our wrapping techniques on a subset of the previous sites, where we manually chose some attributes to extract. Optional attributes are those that occur only on a fraction of the pages in the same Page-Class. When several attributes on a page are optional, their inconsistent occurrence and location are likely to cause extraction rules to fail, so we observe low recall for these tests. Moreover, valid extraction rules cannot be generated when poor HTML formatting affects user selection of data. All in all, the overall results collected on the different web sites support the effectiveness of our tool.

VIII. Related Work

Data extraction for structured web sources has been widely studied, as shown in the various surveys on wrapper generation [7], [8], [9], and the several works on wrapper specification, inference and evaluation of extraction rules (such as [10]). However, our approach focuses on how to contextually specify, create, and execute simple and interrelated crawling and wrapping algorithms, rather independently from the underlying inference mechanisms of extraction rules. A few works [11], [12], [13] have addressed the automatic discovery of content pages and pages containing links to infer a site's hierarchical organization. These approaches mainly aim at finding paths to all generic content pages within a web site, often with some limitations of specific hypotheses to allow automation. In contrast, we aim at semi-automatically gathering pages from a selected class type of interest to a user, with a minimal human effort. The work by [14] partially inspired our URL-REGEX class, as they use URL patterns to specify crawlers in their GoGetIt! system. However, our experiments show that many web sites cannot be crawled in this restrictive way alone. In addition, they adopt a breadth-first strategy to compare the site's pages DOM trees with a provided sample page, while our system works on a small set of sample pages presented to a user for feedback. Sellers A. et al [15] propose a formalism for data extraction by describing and simulating user interactions on dynamic sites. However, the declarative

 ISBN: 978-1-61208-280-6

Table I
CRAWLING RESULTS SUMMARY

Total # of web sites	Total # of pages crawled	Overall F-measure
100	208769	0.99

Table II
WRAPPING RESULTS SUMMARY

Domain	# of web sites	# of pages	# of attributes	# of optional attributes	Overall F-measure
FINANCE	3	610	5	0	1.00
BASKETBALL	4	2310	7	1	0.99
SOCCER	4	1214	5	0	0.99
MOVIES	3	3046	7	6	0.92

specifications are defined by an expert programmer and not derived from an actual user's navigation. Finally, [16] implement a web scale system for data extraction that generates extraction rules on structured *detail* pages in a web site. They apply extensive calculations for clustering pages according to template similarity and rely on several user inputs for annotation and rule learning. Their work is different in scope from ours since we work on synergic crawling and wrapping that can cover various Page-Classes in a web site while deriving information from the user's browsing activity.

IX. CONCLUSION AND FUTURE WORK

We presented in this article a new semi-automatic methodology for synergic crawling and wrapping in the scope of information retrieval. Our contributions can be summarized by: (i) a formalism to specify interleaved crawling programs and wrappers concurrently over structured web sites; (ii) the introduction of the *Sampling Problem*, which illustrates how randomly chosen samples can be biased and negatively impact the inference tasks; (iii) an approach to mitigate the effects of the sampling problem by requiring minimal effort from an inexpert user; (iv) and an experimental evaluation to validate our proposed techniques. Our experiments revealed encouraging results, and can be further improved with the potential inclusion of semantic assertions and a mechanism to deal with any optional attributes with changing location on the page. The quantification of user effort and its variation with the learning implementation parameters is still the subject of an ongoing examination.

REFERENCES

[1] C. Bertoli, V. Crescenzi, and P. Merialdo, "Crawling programs for wrapper-based applications," in *IRI*, 2008, pp. 160–165.

[2] Y. Zhai and B. Liu, "Structured data extraction from the web based on partial tree alignment," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 18, no. 12, pp. 1614–1628, 2006.

[3] S. Chakrabarti, M. Van den Berg, and B. Dom, "Focused crawling: a new approach to topic-specific web resource discovery," *Computer Networks*, vol. 31, no. 11, pp. 1623–1640, 1999.

[4] M. Jamali, H. Sayyadi, B. Hariri, and H. Abolhassani, "A method for focused crawling using combination of link structure and content similarity," in *WI 2006*. IEEE, 2006, pp. 753–756.

[5] J. Madhavan, D. Ko, L. Kot, V. Ganapathy, A. Rasmussen, and A. Halevy, "Google's deep web crawl," *Proc. VLDB Endow.*, vol. 1, no. 2, pp. 1241–1252, Aug. 2008.

[6] V. Crescenzi and P. Merialdo, "Wrapper inference for ambiguous web pages," *Applied Artificial Intelligence*, vol. 22, no. 1&2, pp. 21–52, 2008.

[7] A. H. F. Laender, B. A. Ribeiro-Neto, A. S. da Silva, and J. S. Teixeira, "A brief survey of web data extraction tools," *SIGMOD Record*, vol. 31, no. 2, pp. 84–93, 2002.

[8] S. Flesca, G. Manco, E. Masciari, E. Rende, and A. Tagarelli, "Web wrapper induction: a brief survey," *AI Commun.*, vol. 17, no. 2, pp. 57–61, 2004.

[9] C.-H. Chang, M. Kayed, M. R. Girgis, and K. F. Shaalan, "A survey of web information extraction systems," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1411–1428, 2006.

[10] J. Carme, M. Ceresna, and M. Goebel, "Web wrapper specification using compound filter learning," in *IADIS*, 2006.

[11] V. Crescenzi, P. Merialdo, and P. Missier, "Clustering web pages based on their structure," *Data Knowl. Eng.*, vol. 54, no. 3, pp. 279–299, 2005.

[12] H.-Y. Kao, S.-H. Lin, J.-M. Ho, and M.-S. Chen, "Mining web informative structures and contents based on entropy analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 1, pp. 41–55, 2004.

[13] Z. Liu, W. K. Ng, and E.-P. Lim, "An automated algorithm for extracting website skeleton," in *DASFAA*, 2004, pp. 799–811.

[14] M. L. A. Vidal, A. S. da Silva, E. S. de Moura, and J. M. B. Cavalcanti, "Gogetit!: a tool for generating structure-driven web crawlers," in *WWW*, 2006, pp. 1011–1012.

[15] A. J. Sellers, T. Furche, G. Gottlob, G. Grasso, and C. Schallhart, "Taking the oxpath down the deep web," in *EDBT*, 2011, pp. 542–545.

[16] P. Gulhane, A. Madaan, R. Mehta, J. Ramamirtham, R. Rastogi, S. Satpal, S. Sengamedu, A. Tengli, and C. Tiwari, "Web-scale information extraction with vertex," in *ICDE 2011*. IEEE, 2011, pp. 1209–1220.

 ISBN: 978-1-61208-280-6

Rethinking Traditional Web Interaction

Vincent Balat
Univ Paris Diderot – Sorbonne Paris Cité – PPS, UMR 7126 CNRS, Inria – Paris, France
Email: vincent.balat @ univ-paris-diderot.fr

***Abstract*—Web sites are evolving into ever more complex distributed applications. But current Web programming tools are not fully adapted to this evolution, and force programmers to worry about too many inessential details. We want to define an alternative programming style better fitted to that kind of applications. To do that, we propose an analysis of Web interaction in order to break it down into very elementary notions, based on semantic criteria instead of technological ones. This allows defining a common vernacular language to describe the concepts of current Web programming tools, but also some other new concepts. This results in a significant gain of expressiveness. The understanding and separation of these notions also makes it possible to get strong static guarantees, that can help a lot during the development of complex applications, for example by making impossible the creation of broken links. Most of the ideas we propose have been implemented in the Ocsigen Web programming framework that make possible to write a client-server Web applications as a single program. We will show that the interaction model we propose is fully compatible with this kind of applications.**

***Keywords*—Typing; Web interaction; Functional Web programming; Continuations**

I. INTRODUCTION

Nowadays, Web sites behave more and more like real applications, with a high-level of interactivity on both the server and client sides. For this reason, they deserve well-designed programming tools, with features like high-level code structuring and static typing. These tools must take into account the specificities of that kind of application. One of these specificities is the division of the interface into *pages*, connected to each other by links. These pages are usually associated to *URLs* which one can bookmark. It is also possible to turn back to one page using the *back button*. This makes the dynamics of the interface completely different from a regular application. Another specificity is that this kind of applications is highly dependent on standards as they will be executed on various platforms.

Web programming covers a wide range of fields, from database to networking. The ambition of this paper is not to address them all, nor to deal with the full generality of *service oriented computing*. We concentrate on what we will call *Web interaction*; that is, the interaction between a user and a Web application, through a browser interface. We place ourselves in the context of writing such an application, that communicates with one or several servers, and with the ability to make part of the computation in the browser. A few similar Web interaction systems have already been described before (for example Links [1], or Hop [2]). The goal of this paper is mainly to focus on one feature that has been very rarely addressed before, namely: *service identification*, that is, how the service to handle a request is chosen. We want to show that a good service identification mechanism can help programmers a lot, and that the concepts we present here allow to take into account very concrete needs of Web developpers that are usually not addressed by more theoretical works.

We want to make a first step towards a formalization of this interaction with a twofold goal. First, we want to increase the expressiveness of Web frameworks. Second, we want to improve the reliability of Web applications by using well defined concepts and static validation of the code.

The concepts we present here have been implemented in the Ocsigen Web programming framework [3], [4], [5] (Eliom project). It allows to program fully in OCaml both the server and client parts of a Web application, with a consistent abstraction of concepts. A compiler to Javascript is used to run the client parts in the browser [6], [7].

Work partially supported by the French national research agency (ANR), PWD project, grant ANR-09-EMER-009-01, and performed at the IRILL center for Free Software Research and Innovation in Paris, France

A. A common vernacular language for Web interaction

Web development is highly constrained by technologies. First, it relies on the HTTP protocol, which is non-connected and stateless. Then, Web applications must be executable in various browsers that implement more or less accurately common standards and recommendations.

We want to detach ourselves as much as possible from these constraints and think about how we would like to program Web interaction. Obviously this is also a question of taste. But rather than proposing yet another programming model from scratch, we start by analyzing common Web programming practices, in order to understand the notions they use. Then we decompose them in very elementary notions that can be used to describe the features of Web programming tools, from PHP to JSP or Miscrosoft.NET Web Forms, etc. We will observe that current frameworks impose too many artificial restrictions. Ideally, we would like to give a generic language, flexible enough to describe all possible behaviours, without imposing any artificial restriction due to one technology.

We place ourselves at a semantic level rather than at a technical one. Moving away from technical details will allow to increase the *expressiveness* of Web programming frameworks. In the domain of programming languages, high-level concepts have been introduced over the years, for example genericity, inductive types, late binding, closures. They make easier the implementation of some complex behaviours. We want to do the same for the Web. For example the notion of *"sending a cookie"* benefits from being abstracted to a more semantic notion like *"opening a session"* (which is already often the case today). Also it is not really important for the programmer to know how URLs are formed. What matters is the service we want to speak about (and optionally the parameters we want to send it).

This abstraction from technology allows two things:

- First, it increases the expressiveness of the language by introducing specific concepts closer to the behaviours we want to describe (and irrespective of the way they are implemented). From a practical point of view, this allows to implement complex behaviours in very few lines of code.
- Having well-designed dedicated concepts also allows to avoid wrong behaviours. We forbid unsafe technical possibilities either by making them inexpressible, or by static checking.

B. Improving the reliability of Web applications

As Web sites are currently evolving very quickly into complex distributed applications, the use of strongly and statically typed programming languages for the Web becomes more and more helpful. Using scripting languages was acceptable when there was very little dynamic behaviour in Web pages, but current Web sites written with such languages are proving to be very difficult to evolve and maintain. Some frameworks are counterbalancing their weaknesses by doing a lot of automatic code generation (for example [8]). But this does not really improve the *safety* of programs. In the current state of knowledge, we are able to do much better, and Web programming must benefit from this.

 ISBN: 978-1-61208-280-6

Static validation of pages: One example where static typing revolutionizes Web programming concerns the validation of pages. Respecting W3C recommendations is a necessity to ensure portability and accessibility of Web sites. The novelty is that there now exist typing systems sophisticated enough to statically ensure a page's validity [9], [10], [11] We do not mean checking the validity of pages once generated, but really to be sure that the program that builds the XML data will always generate something valid, even in the most particular cases.

For example, even if a programmer has checked all the pages of his site in a validator, is he sure that the HTML table he creates dynamically will never be empty (which is forbidden)? What if for some reason there is no data? He must be very conscientious to think about all these cases. It is most likely that the evolutions of the program will break the validity of pages. In most cases, problems are discovered much later, by users.

In lots of cases, such errors will even make the generated output unusable, for example for XML data intended to be processed automatically. The best means to be sure that this situation will never happen is to use a typing system that will prevent one from putting the service on-line if there is the slightest risk for something wrong to be generated.

For people not accustomed to such strong typing systems, this may seem to impose too much of a constraint to programmers. Indeed, it increases a bit the initial implementation time (by forcing to take into account all cases). But it also saves such a huge amount of debugging time, that the use of such typing systems really deserves to be generalized. For now, these typing systems for XML are used in very few cases of Web services, and we are not aware of any major Web programming framework. Our experience is that it is not difficult to use once one get used to the main rules of HTML grammar, if error messages are clear enough.

Validity of Web interaction: Static checking and abstraction of concepts can also benefit in many other ways to Web programming, and especially to Web interaction. Here are a few examples: In a link, do the types (and names) of parameters match the types expected by the service it points to? Does a form match the service it points to? Do we have broken links?

It is not so difficult to have these guarantees, even if almost no Web programming framework are doing so now. All what is needed is a programming language *expressive* enough (in the sense we explained above).

Improving the ergonomics of Web sites: Lots of Web developers are doing implementation errors resulting in reduced ease of use (wrong use of sessions or GET and POST parameters, etc.). Take as example a famous real estate Web site that allows to browse through the results of a search; but if someone sets a bookmark on one of the result pages, he never goes back to the same page, because the URL does not refer to the advertisement itself, but to the rank in the search. We will see that a good understanding of concepts can avoid such common errors.

C. Overview of the paper

Sections II and III are devoted to the definition of our vernacular language for describing the services provided by a Web application. Section II explains the advantage of using an abstract notion of service instead of traditional page-based programming and string URLs. Section III presents a new service identification and selection method. It shows how powerful this notion of service can be made, by separating it into several kinds. This results in a very new programming style for Web interaction.

II. ABSTRACTING SERVICES

As explained above, we want to formalize Web interaction, that is, the behaviour of a Web application in reaction to the actions of the user. What happens when somebody clicks on a link or submits a form? A click often means that the user is requesting a new document: for example a new page that will replace the current one (or one part of it). But it can also cause some actions to take place on the server or the client. Let us enumerate the different kinds of reactions. A click (or a key strike) from the user may have the following main effects:

1) Modifying the application interface. That is, changing the page displayed by the browser (or one part of the page), or opening a new window or tab with a new page,
2) Changing the URL displayed by the browser (protocol, server name, path, parameters, etc.),
3) Doing some other action, like the modification of a state (for example changing some database values),
4) Sending hidden data (like form data, or files),
5) Getting some data to be saved on the user's hard disk.

Two important things to notice are that each of these items is optional, and may either involve a distant server, or be processed locally (by the browser).

This decomposition is important, as a formalization of Web interaction should not omit any of these items in order not to restrict the freedom of the programmer. All these items are described semantically, not technically.

A. The role of URLs

The item "Changing the URL" above is a really significant one and is one key to understand the behaviour of Web applications. This section is devoted to the understanding of that notion. URLs are entry points to the Web site. Changing the URL semantically means: giving the possibility to the user to turn back to this point of interaction later, for example through bookmarks.

Note that, unlike many Web sites, a good practice is to keep the URL as readable as possible, because it is an information visible to users that may be typed manually.

1) Forgetting technical details about URLs: The syntax of URLs is described by the Internet standard STD 66 and RFC 3986 and is summarized (a bit simplified) here:
`scheme://user:pwd@host:port/path?query#fragment`

The path traditionally describes a file in the tree structure of a file system. But this view is too restrictive. Actually, the path describes the hierarchical part of the URL. This is a way to divide a Web site into several sections and subsections.

The query string syntax is commonly organized as a sequence of '*key=value*' pairs separated by a semicolon or an ampersand, e.g., `key1=value1&key2=value2&key3=value3`. This is the part of the URL that is not hierarchical.

To a first approximation, the path corresponds to the service to be executed, and the query to parameters for this service. But Web frameworks are sometimes taking a part of the path as parameters. On the contrary, part of the query, or even of the host, may be used to determine the service to call. This will be discussed later in more detail.

The *fragment* part of the URL only concerns the browser and is not sent to the server.

The item "Changing the URL" is then to be decomposed semantically into these sub-tasks:

1) Changing the protocol to use,
2) Changing the server (and port) to which the request must be made,
3) Choosing a hierarchical position (path) in the Web site structure, and specifying non hierarchical information (query) about the page,
4) And optionally: telling who the user is (credentials) and the fragment of the page he wants to display.

 ISBN: 978-1-61208-280-6

2) URL change and service calls: There are two methods to send form data using a browser: either in the URL (GET method) or in the body of the HTTP request (POST method). Even if they are technical variants of the same concept (a function call), their semantics are very different with respect to Web interaction. Having parameters in the URL allows to turn back to the same document later, whereas putting them in the request allows to send one-shot data to a service (for example because they will cause an action to occur).

We propose to focus on this semantical difference rather than on the way it is implemented. Instead of speaking about POST or GET parameters, we prefer the orthogonal notions of *service calls* and *URL change*. It is particularly important to forget the technical details if we want to keep the symmetry between server and client side services. Calling a local (javascript for example) function is similar to sending POST data to a server, if it does not explicitly change the URL displayed by the browser.

Semantically speaking, in modern Web programming tools, changing the URL has no relation with *calling a service*. It is possible to call a service without changing the URL (because it is a local service, or because the call uses POST parameters). On the contrary, changing the URL may be done without calling a service. There is only one reason to change the URL: give the user a new entry point to the Web site, to which he can come back when he wants to ask the same service once again, for example by saving it in a bookmark.

B. Services as first class values

The first main principle on which is based our work is: *consider services as first class values,* exactly as functional languages consider functions as first class valuesThat is: we want to manipulate services as abstract data (that can for example be given as parameter to a function). This has several advantages, among which:

- The programmer does not need to build the syntax of URLs himself. Thus, it is really easy to switch between a local service and a distant one.
- All the information about the service is taken automatically from the data structure representing the service, including the path to which the service is attached and parameter names. This has a very great consequence: if the programmer changes the URL of a service, even the name of one of its parameters, he does not need to change any link or form towards this service, as they are all built automatically. This means that links will never be broken, and parameter names will always be correct (at least for internal services, i.e., services belonging to the Web site).

Some recent frameworks already have an abstraction of the notion of service. We want to show how to take the full benefit of it. Our notion of service must be powerful enough to take into account all the possibilities described above, but without relying on their technical implementation.

A service is some function taking parameters and returning some data, with possibly some side effects (remote function calls). The server is a provider of *services*. Client side function calls can also be seen as calls to certain services. The place where services take place is not so significant. This allows to consider a Web site with two versions of some services, one on server side, the other on client side, depending on the availability of some resources (network connection, or browser plug-ins for example).

The language must provide some way to define these services, either using a specific keyword or just through a function call.

Once we have this notion, we can completely forget the old "page-based" view of the Web where one URL was supposed to be associated to one file on the hard disk. Thus, it is possible to gain a lot of freedom in the organization and modularity of the code, and also, as we will see later, in the way services are associated to URLs. One of the goals of next section is precisely to discuss service identification and selection, that is, how services are chosen by the server from the hierarchical and non-hierarchical parts of the URL, and hidden parameters.

III. A TAXONOMY OF SERVICES

A. Values returned by services

A first classification of services may be made according to the results they send. In almost all Web programming tools, services send HTML data, written as a string of characters. But as we've seen before, it is much more interesting to build the output as a tree to enable static type checking. To keep full generality, we will consider that a service constructs a result of any type, that is then sent, possibly after some kind of serialization, to the browser which requested it. It is important to give to the programmer the choice of the kind of service he wants.

A reflection on return types of services will provide once again a gain of expressiveness. Besides plain text or typed HTML trees, a service may create for example a redirection. One can also consider using a service to send a file. It is also important to give the possibility to services to choose themselves what they want to send. For example, some service may send a file if it exists, or an HTML page with an error message on the other case. The document is sent together with its *content type*, telling the browser how to display it (it is a dynamic type). But in other cases, for example when a service implements a function to be called from the client side part of the program, one probably want the type of the result to be known statically.

We also introduce a new kind of output called *actions*. Basically sending an action means "no output at all". But the service may perform some action as side effect, like modifying a database, or connecting a user (opening a new session). From a technical point of view, actions implemented server side are usually sending a 204 (No content) HTTP status code. Client side actions are just procedures. We will see some examples of use of actions and how to refine the concept in Section III-D2.

B. Dynamic services, or continuation-based Web programming

1) Dynamic services: Modern Web frameworks propose various solutions to get rid of the lack of flexibility induced by a one-to-one mapping between one URL and one service. But almost none of them take the full benefit of this, and especially of one very powerful consequence: the possibility to dynamically create new services.

For example, if we want to add a feature to a Web site, or even if we occasionally want to create a service depending on previous interaction with one user. For example, if one user wants to book a plane ticket, the system will look in a database for available planes and dynamically create the services corresponding to booking each of them. Then it displays the list of tickets, with, on each of them, a link towards one of these dynamic services. Thus, we will be sure that the user will book the ticket he expects, even if he duplicates his browser window or uses the back button. This behaviour is really simple to implement with dynamic services and rather tricky with traditional Web programming. Witness the huge number of Web sites which do not implement this correctly.

If we want to implement such behaviour without dynamic services, we will need to save somewhere all the data the service depends on. One possibility is to put all this data in the link, as parameters, or in hidden form data. Another possibility, for example if the amount of data is prohibitive, is to save it on the server (for example in a database table) and send only the key in the link.

With dynamic service creation, all this contextual data is recorded automatically in the *environment* of the closure implementing the service. This closure is created dynamically according to some dynamic data (recording the past of the interaction with the user). It requires a functional language to be implemented easily.

 ISBN: 978-1-61208-280-6

2) Continuations: This feature is equivalent to what is known as *continuation-based Web programming*. This technique was first described independently by Christian Queinnec [12], [13], [14], John Hughes [16] and Paul Graham [15].

The use of dynamic services is a huge step in the understanding of Web interaction, and an huge gain of expressiveness. Up until now, very few tools have used these ideas. None of the most widely used Web programming frameworks implement them, but they are used for example in Seaside [17], PLT Scheme [18], Hop [2], Links [1], and obviously Ocsigen.

The cost (in terms of memory or disk space consumption) is about the same as with usual Web programming: no copy of the stack, one instance of the data, plus one pointer to the code of the function.

C. Finding the right service

The very few experimental frameworks which are proposing some kind of dynamic services impose usually too much rigidity in the way they handle URLs. This section is devoted to showing how it is possible to define a notion of service identification that keeps all the possibilities described in Section II.

The important thing to take care of is: how to do the association between a request and a service? For example if the service is associated to an URL, where, in this URL, is the service to be called encoded?

To make this as powerful as possible, we propose to delegate to the server the task of decoding and verifying parameters, which is traditionally done by the service itself. This has the obvious advantage of reducing a lot the work of the service programmer. Another benefit is that the choice of the service to be called can depend on parameters.

Let us first speak about distant (server side) bookmarkable services, i.e., services called by sending a GET request to a server. We will speak later about client side services, and hidden services.

1) Hierarchical services: One obvious way to associate a service to an URL is by looking at the path (or one part of it). We will call these services *hierarchical services*. These kinds of services are usually the main entry points of a Web site. They may take parameters, in the *query* part of the URL, or in the path. One way to distinguish between several hierarchical services registered on the same path is to look at parameters. For example the first registered service whose expected parameters exactly match the URL will answer.

2) Coservices: Most of the time one probably wants dynamic services to share their path with a hierarchical service, at least those which last for only a short time (result of a search for example). Also one may want two services to share the same hierarchical position on the Web site.

We will call *coservices* services that are not directly associated to a path, but to a special parameter. This is one of the main original features of our service identification mechanism and this has a huge impact on expressiveness, as we will see on example in Section III-D2. From a semantic point of view, the difference is that hierarchical services are the entry points of the site. They must last forever, whereas coservices may have a timeout, and one probably want to use the associated main service as fallback when the coservice has expired.

We will distinguish between *named coservices* and *anonymous coservices*, the difference being the value of the special parameter. Named coservices have a fixed parameter value (the name of the coservice), whereas this value is generated automatically for anonymous coservice.

Like all other services, coservices may take parameters, that will be added to the URL. There must be a way to distinguish between parameters for this coservice and parameters of the original service. This can be done by adding automatically a prefix to coservice parameters.

3) Attached and non-attached coservices: We will also distinguish between coservices *attached* to a path and *non-attached* coservices. The key for finding an attached coservice is the path completed by a special parameter, whereas non-attached coservices are associated to a parameter, whatever the path in the URL. This feature is not so common and we will see in Section III-D2 how powerful it is.

4) Distant hidden services: A distant service is said to be *hidden* when it depends on POST data sent by the browser. If the user comes back later, for example after having made a bookmark, it will not answer again, but another service, not hidden, will take charge of the request. We will speak about *bookmarkable* services, for services that are not hidden.

Hidden services may induce an URL change. Actually, we can make exactly the same distinction as for bookmarkable services: there are hierarchical hidden services (attached to a path), hidden attached coservices (attached to a path, and a special POST parameter), and hidden non-attached coservices (called by a special POST parameter).

It is important to allow the creation of hidden hierarchical services or coservices only if there is a bookmarkable (co)service registered at the same path. This service will act as a fallback when the user comes back to the URL without POST parameters. This is done by specifying the fallback instead of the path when creating a hidden service. It is a good idea to do the same for bookmarkable coservices.

Registering a hidden service on top of a bookmarkable service with parameters allows to have both GET and POST parameters for the same service. But bear in mind that their roles are very different.

5) Client side services: Client side service calls have the same status as hidden service calls. In a framework that allows to program both the server and client sides using the same language, we would like to see local function calls as non-attached (hidden) coservices. Hierarchical hidden services and (hidden) attached coservices correspond to local functions that would change the URL, without making any request to the server.

D. Taxonomy of services

1) Summary of service kinds: Figure 1 summarizes the full taxonomy of services we propose. This set is obviously *complete* with respect to technical possibilities (as traditional services are part of the table). It is powerful enough for describing in very few lines of code lots of features we want for Web sites, and does not induce any limitations with respect to the needs of Web developers. Current Web programming frameworks usually implement a small subset of these possibilities. For example “page-based” Web programming (like PHP or CGI scripts) does not allow for non-attached coservices at all. Even among “non-page-based” tools, very few allow for dynamic (anonymous) coservice creation. To our knowledge, none (but Ocsigen) is implementing actions on non-attached services as primary notions (even if all the notions can obviously be simulated).

2) Example cases: We have already seen some examples of dynamic service creation: if a user creates a blog in a subdirectory of his or her personal site, one possibility is to add dynamically a hierarchical service to the right path (and it must be recreated every time the serser is relaunched). If we want to display the result of a search, for example plane ticket booking, we will create dynamically a new anonymous coservice (hidden or not), probably with a timeout. Without dynamic services, we would need to save manually the search keyword or the result list in a table.

Coservices are not always dynamic. Suppose we want a link towards the main page of the site, that will close the session. We will use a named hidden attached coservice (named, so that the coservice key is always the same).

We will now give an example where non-attached hidden coservices allow to reduce significantly the number of lines of code. Consider a site with several pages. Each page has a connected version and a non-connected version, and we want a connection box on

 ISBN: 978-1-61208-280-6

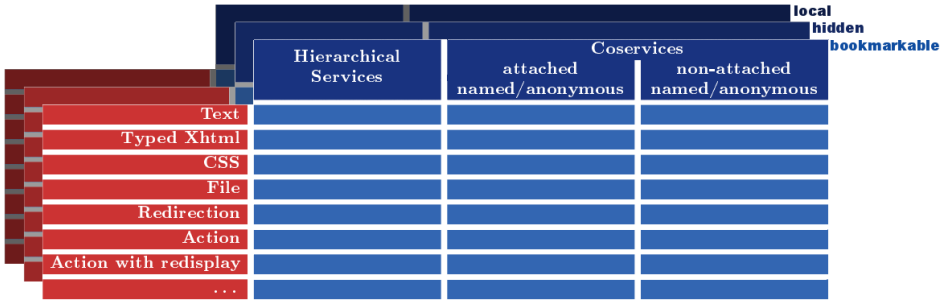


Figure 1: Full taxonomy of services.

each non-connected page. But we don't want the connection box to change the URL. We just want to log in and stay on the same URL, in connected version. Without non-attached services (and thus with almost all Web programming tools), we need to create a version with POST parameters of each of our hierarchical services to take into account the fact that each URL may be called with user credentials as POST parameters.

Using our set of services, we just need to define only one non-attached (hidden) coservice for the connection. At first sight, that service only performs an action (as defined in Section III-A): saving user information in a session table. But we probably want to return a new page (connected version of the same page). This can be done easily by returning a redirection to the same URL. Another solution if we don't want to pay the cost of a redirection, is to define a new kind of output: *"action with redisplay"* that will perform the action, then make an internal (server side) request as if the browser had done the redirection. The solution with redirection has one advantage: the browser won't try to resend POST data if the user reloads the page.

Now say for example that we want to implement a wiki, where each editable box may occur on several pages. Clicking on the *edit* button goes to a page with an edit form, and submitting the form must turn back to the original page. One dirty solution would be to send the original URL as hidden parameter in the edit link. But there is now a simpler solution: just do not change the path. The edit form is just a page registered on a non-attached service.

Our reflexion on services, also allows to express clearly a solution to the real estate site described in Section I-B. Use (for example) one bookmarkable hierarchical service for displaying one piece of advertisement, with additional (hidden or not) parameters to recall the information about the search.

3) Expressiveness: The understanding of these notions and their division into very elementary ones induces a significant gain in expressiveness. This is particularly true for actions with redisplay. They are very particular service return values and seem to be closely related to non-attached coservices at first sight. But the separation of these two concepts introduces a new symmetry to the table, with new cells corresponding to very useful possibilities (see the example of the wiki above). It is noteworthy that all cells introduced in this table have shown to be useful in concrete cases.

IV. Conclusion

A. Related work

A lot of modern Web programming frameworks (for example GWT or Jif/Sif [19]) are trying to propose integrated and high level solutions to make easier the development of Web application. They often provide some abstraction of concepts, but most of them preserve some historical habits related to technical constraints. It is impossible to make a full review of such tools, as there are numerous. We will concentrate on the main novel features presented here. One can try to make a classification of existing Web frameworks with respect to the way they do service identification.

The old method is what we called "page-based Web programming", where one path corresponds to one file. Modern tools are all more flexible and make service identification and selection independent of the physical organization of components in the Web server (for example JSP assigns an URL to a service from a configuration file). But very few belong to the third group, that allows dynamic services. Among them: Seaside [17], Links [1] and Hop [2], Wash/CGI [20]. Their service identification models are more basic, and they don't have a native notion of coservice. Some of them are using an abstraction of forms [20], [21] that is fully compatible with our model.

There have been few attempts to formalize Web interaction. The most closely related work is by Paul T. Graunke, Robert Bruce Findler, Shriram Krishnamurthi and Matthias Felleisen [22], [23]. Their work is more formal but does not take into account all the practical cases we speak about. In particular their service model is much simpler and does not fully take into account the significance of URLs. Peter Thiemann [20] uses monads to create HTML pages, which makes possible an original and interesting way of handling the typing of forms, using Haskell's type system.

We think our approach is compatible with more data driven approaches [8], component based interfaces [24], or even code generation techniques. One interesting work would be to see how they can be mixed together.

B. Evolution of technologies and standards

This reflection about Web programming techniques has shown that Web technologies suffer from some limitations that slow down the evolution towards really dynamic applications. Here are a few examples:

- As mentioned above, the format of page parameters and the way browsers send them from form data does not allow for sophisticated parameters types.
- (X)HTML forms cannot mix GET and POST methods. It is possible to send URLs parameters in the `action` attribute of a form that is using the POST method, but it is not possible to take them from the form itself. This would open many new possibilities.
- A link from HTTP towards the same site in HTTPS is always absolute. This breaks the discipline we have to use only relative links (for example to behave correctly behind a reverse proxy). We have the same problem with redirections, which have to be absolute URLs according to the protocol.
- There is no means to send POST data through a link, and it is difficult to disguise a form into a link. Links and forms should probably be unified into one notion, that would allow to make a (POST or GET) request from a click on any part of the page. This limitation is not really significant if we have a client side program that does the requests itself when we click on a page element.
- Having the ability to put several `id` attributes for one tag would be very useful for automatically generated dynamic pages.

 ISBN: 978-1-61208-280-6

- Probably one of the main barriers to the evolution of the Web today is the impossibility to run fast code on the browser (without plug-ins), even with recent implementations of Javascript. When thinking about a Web application as a complex application distributed between a server and a client, we would often like to perform computationally intensive parts of the execution on the client, which is not feasible for now. We want to make some experiments with Google Native Client [25].

C. Concluding words and future works

This papers presents a new programming style for Web interaction which simplifies a lot the programming work and reduces the possibilities of semantical errors and bad practices. The principles we advocate are summarized here:

1) Services as first class values
2) Decoding and verification of parameters done by the server
3) Dynamic creation of services
4) Full taxonomy of services for precise service identification
5) Same language on server and client sides
6) Symmetry between local and distant services

One of the main novel feature is the powerful service identification mechanism performed automatically by the server. It introduces the notion of coservice which make the programming of sophisticated Web interaction very easy.

Beyond just presenting a new Web programming model, this paper defines a new vocabulary for describing the behaviour of Web sites, on a semantic basis. It is a first step towards a formalization of Web interaction. We started from an analysis of existing Web sites and we extracted from this observation the underlying concepts, trying to move away as much as possible from non-essential technical details. This allowed a better understanding of the important notions but above all to bring to light some new concepts that were hidden by technical details or historical habits. The main feature that allowed this is the introduction of dynamic services, and also forgetting the traditional page-based Web programming. There exist very few frameworks with these features, and none is going as far as we do, especially in the management of URLs.

Besides the gain in expressiveness, we put the focus on reliability. This is made necessary by the growing complexity of Web applications. The concepts we propose allow for very strong static guarantees, like the absence of broken links. But more static checks can be done, for example the verification of adequacy of links and forms to the service they lead to. These static guarantees have not been developed here because of space limitation. They are summarized by the following additional principles:

7) Static type checking of generated data
8) Static type checking of links and forms

This paper does not present an abstract piece of work: all the concepts we present have been inspired by our experience in programming concrete Web sites, and have been implemented. Please refer to Ocsigen's manual [26] and source code for information about the implementation. Some implementation details may also be found in [27] (describing an old version of Ocsigen that was using a more basic model of services). Ocsigen is now used in industry (for example BeSport [29], Pumgrana [28]). These concrete experiences showed that the programming style we propose is very convenient for Web programmers and reduces a lot the work to be done on Web interaction.

This paper is not a full presentation of Ocsigen. Many aspects have been hidden, and especially how we program the client side part of the application [6], [7], [30] using the same language, and with the same strong static guarantees. As we have seen, our notions of services also apply to client side functions. Obviously we are using the same typing system for services but also for HTML. It is not easy to guarantee that a page will remain valid if it can evolve over time [31]. We did not show how the server can send data to the client at any time, or even call a function on client side.

REFERENCES

[1] E. Cooper, S. Lindley, P. Wadler, and J. Yallop, "Links: Web programming without tiers," in In 5th International Symposium on Formal Methods for Components and Objects (FMCO). Springer-Verlag, 2006.

[2] M. Serrano, E. Gallesio, and F. Loitsch, "Hop, a language for programming the web 2.0," in Dynamic Languages Symposium, Oct. 2006.

[3] The Ocsigen project, http://www.ocsigen.org [retrieved: 03, 2013].

[4] V. Balat, J. Vouillon, and B. Yakobowski, "Experience report: ocsigen, a web programming framework," in ICFP '09: Proceedings of the 14th ACM SIGPLAN international conference on Functional programming.

[5] V. Balat, P. Chambart, and G. Henry, "Client-server Web applications with Ocsigen," in WWW2012 dev track proceedings

[6] B. Canou, E. Chailloux, and J. Vouillon, "How to Run your Favorite Language in Web Browsers," in WWW2012 dev track proceedings

[7] J. Vouillon and V. Balat, "From bytecode to javascript: the js_of_ocaml compiler," Journal of Software: Practice and Experience, 2013.

[8] Ruby on rails, http://www.rubyonrails.com/ [retrieved: 03, 2013]

[9] A. Frisch, "Ocaml + xduce," in Proceedings of the international conference on Functional programming (ICFP). ACM, 2006

[10] V. Benzaken, G. Castagna, and A. Frisch, "CDuce: An XML-centric general-purpose language," in Proceedings of the International Conference on Functional Programming (ICFP), 2003

[11] H. Hosoya and B. C. Pierce, "XDuce: A statically typed XML processing language," ACM Transactions on Internet Technology, vol. 3, no. 2, May 2003.

[12] C. Queinnec, "The influence of browsers on evaluators or, continuations to program web servers," in International conference on Functional programming (ICFP), 2000

[13] C. Queinnec, "Continuations and web servers," Higher-Order and Symbolic Computation, Dec. 2004.

[14] C. Queinnec, "Inverting back the inversion of control or, continuations versus page-centric programming," ACM SIGPLAN Notices, vol. 38, no. 2, Feb. 2003.

[15] P. Graham, "Beating the averages" http://www.paulgraham.com/avg.html [retrieved: 03, 2013].

[16] J. Hughes, "Generalising monads to arrows," Science of Computer Programming, vol. 37, no. 1–3 2000.

[17] S. Ducasse, A. Lienhard, and L. Renggli, "Seaside – a multiple control flow web application framework," in Proceedings of ESUG Research Track 2004, 2004.

[18] S. Krishnamurthi, P. W. Hopkins, J. Mccarthy, P. T. Graunke, G. Pettyjohn, and M. Felleisen, "Implementation and use of the plt scheme web server," in Higher-Order and Symbolic Computation, 2007.

[19] S. Chong, J. Liu, A. C. Myers, X. Qi, K. Vikram, L. Zheng, and X. Zheng, "Secure web applications via automatic partitioning," SIGOPS Oper. Syst. Rev., vol. 41, no. 6, 2007.

[20] P. Thiemann, "Wash/cgi: Server-side web scripting with sessions and typed, compositional forms," in Practical Aspects of Declarative Languages (PADL'02), 2002

[21] E. Cooper, S. Lindley, P. Wadler, and J. Yallop, "The essence of form abstraction," in Sixth Asian Symposium on Programming Languages and Systems, 2008.

[22] P. T. Graunke, R. B. Findler, S. Krishnamurthi, and M. Felleisen, "Modeling web interactions," in European Symposium on Programming (ESOP), April 2003

[23] S. Krishnamurthi, R. B. Findler, P. Graunke, and M. Felleisen, "Modeling web interactions and errors," in In Interactive Computation: The New Paradigm. Springer Verlag, 2006.

[24] J. Yu, B. Benatallah, R. Saint-Paul, F. Casati, F. Daniel, and M. Matera, "A framework for rapid integration of presentation components," in WWW '07: Proceedings of the 16th international conference on World Wide Web. ACM, 2007.

[25] "Google native client," http://code.google.com/p/nativeclient/ [retrieved: 03, 2013].

[26] V. Balat, "Eliom programmer's guide," Technical report, Laboratoire PPS, CNRS, université Paris-Diderot, Tech. Rep., 2007

[27] V. Balat, "Ocsigen: Typing web interaction with objective caml," in ML'06: Proceedings of the 2006 ACM SIGPLAN workshop on ML

[28] "Pumgrana," http://www.pumgrana.com/ [retrieved: 03, 2013].

[29] "BeSport," http://www.besport.com [retrieved: 03, 2013].

[30] B. Canou, V. Balat, and E. Chailloux, "O'browser: objective caml on browsers," in ML '08: Proceedings of the 2008 ACM SIGPLAN workshop on ML.

[31] B. Canou, V. Balat, and E. Chailloux, "A declarative-friendly api for web document manipulation," in International Symposium on Practical Aspects of Declarative Languages (PADL'13). Springer-Verlag, January 2013.

 ISBN: 978-1-61208-280-6

Guided-Based Usability Evaluation On Mobile Websites

Bahadır Dündar
Software Testing and Quality Evaluation Center
TUBITAK
Gebze-Kocaeli, TURKEY
bahadir.dundar@tubitak.gov.tr

Nejat Yumusak
Computer Engineering Department
Sakarya University
Sakarya, TURKEY
n.yumusak@sakarya.edu.tr

Samet Arsoy
Computer Engineering Department
Yildiz Technical University
Istanbul, TURKEY
samet.arsoy@gmail.com

***Abstract*— We live in a mobile world and usability is an important issue for mobile. Considering two mobile websites, how can you tell one of them is more usable and user friendly than the other one? This paper shows how to make a usability evaluation on mobile websites. It covers many guidelines and researches offered to designers to make user friendly interfaces. In this study, we introduce a usability evaluation tool that provides an environment for usability experts and describe how we use it in our Guided-Based Usability Evaluation Model (GBUEM). This study is aimed to contribute to mobile usability area, which is an important and hot topic.**

Keywords-usability; mobile; evaluation; usability guidelines

I. INTRODUCTION

In today's world, the use of mobile website is increasing day by day and internet access from mobile devices is gaining importance. A substantial part of Web users connect internet from mobile devices. Each year there is a huge growth in the mobile share of web traffic across the world. In the fourth quarter of 2012, 23.14 % of total website traffic was generated via mobile devices, while the same number was a mere 12.59% in 2011 [1]. Mobile devices have advantages compared to desktop computers from the point of accessibility and portability.

Most of the operations that can be done with regular websites can also be done via mobile websites, which are specifically designed for mobile devices. Success rate of mobile website usability is 64%, while the same number is 58% for regular websites, which are designed for full-screen desktop computers [2]. It leads us to the fact that website designers should also focus on building mobile websites besides regular websites.

One of the most important problems of mobile Web is the usability problem. The mobile Web provides very different set of challenges, such as limited bandwith, limited input capabilities, small screen size, and no flash script capabilities [3]. Because of this, there can be many usability problems for the end users. Users can revisit the websites if they are satisfied in performing their tasks. To this end, increasing mobile usability standards is important for helping users to enhance their web surfing experience in mobile websites. Usability problems need to be investigated and solved in order to make end user more satisfied. In this context, we argue that evaluation of the usability of mobile websites has significance.

This paper includes following sections. Section II presents the definition of mobile usability and background of our study. Section III presents the Guided-Based Usability Evaluation on Mobile Websites. Section IV presents the practice of proposal methodology. Finally, section V presents the future work and conclusion of our study.

II. BACKGROUND FOR MOBILE USABILITY

A. Usability Definition

There are several usability definitions. ISO 9241-11 gives the conventional usability definition. It defines usability as "the extent to which a product can be used with effectiveness (number/percentage of completed tasks within allotted time, number of errors), efficiency (time to complete a task) and satisfaction (subjective user attitude) in a specified context of use" [4].

On the other hand, usability is defined by five quality components by Nielsen Norman Group, which is the leading research group in usability [5].

- Learnability: how easy users perform their basic tasks in the first place?
- Efficiency: how quickly can users perform a task?
- Memorability: how easy is to reuse the system after a break?
- Errors: How many errors do users make and how serious are these errors using the interface?
- Satisfaction: How do users like the interface?

B. Mobile Usability Background

Usability is a key concept in mobile world because in mobile world tasks are time sensitive. Mobile websites should provide user satisfaction. If a website is not user friendly, users may leave the website page immediately. On average, the user spends less than a minute on a web page [6]. This is a very little time for websites. In order to increase

 ISBN: 978-1-61208-280-6

this number, websites should be designed based on user perspective.

Mobile devices, which are designed for a specific function, can perform additional tasks with some limited processing capabilities, wireless network connection and memory [7]. Mobile devices have more limitations compared to desktop computers. These limitations can cause many usability problems. Mobile web usability success rate of today is almost the same as the desktop usability success rate of 1999. Current desktop usability success rate is 84%, and unless its mobile counterpart starts improving rapidly, it will not reach that level until 2026 [2]. This demonstrates that there is room for research and development in mobile usability.

It is about twice as hard to understand complicated content when reading in touch screen based mobile phones compared to desktop computers. In this study, the only parameter that is accounted for is the screen size, which is the reason for lower comprehension score in mobile devices [8]. According to this research, it can be concluded that a separate mobile version of website should be designed to have a better understanding of the content for the user. The website should automatically detect the mobile device of the user, and divert it to the mobile website [9].

Traditional usability methods that are employed for desktop computer environments, are not directly applicable for mobile environments due to certain mobile specifications. New approaches and methodologies should be considered in the context of mobile. Mobile devices and mobile technologies change rapidly therefore usability evaluation methods and guidelines should be revised.

There are other studies on mobile web environment, which investigates the environmental effects on mobile web usability, mobile web usability documentation, remote usability testing, lab and environmental mobile usability testing. Nielsen Norman Group has also two reports on mobile web usability [9-12]. Contribution of our study is defining a practical way and presenting an environment for experts to evaluate a mobile website according to researches and guidelines.

III. GUIDED-BASED USABILITY EVALUATION

In this study, we show how GBUEM can be applied on mobile websites in a practical way. In this context, guidelines, which make recommendation for mobile websites in terms of usability, are analyzed. In the analysis process, these guidelines are categorized and list of rule is defined. Fig. 1 demonstrates the flow chart for how GBUEM is performed.

A. Usability Standards and Guidelines

ISO-9241-151 Guidance on World Wide Web User Interfaces provides guidance on the human-centred software design for Web user interfaces with the aim of increasing usability.

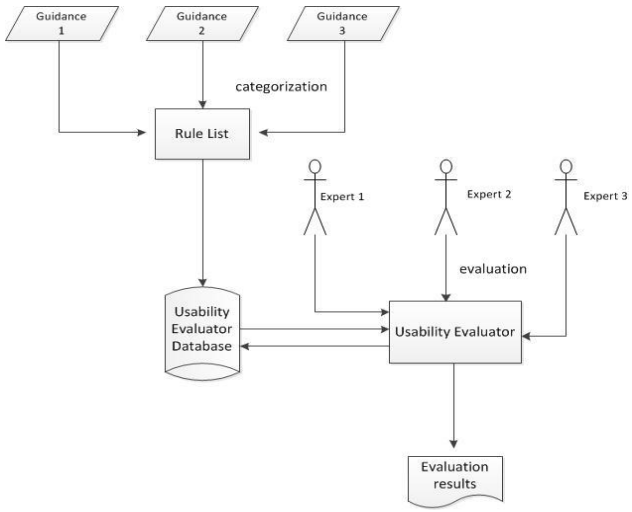


Figure 1. Guided-Based Usability Evaluation Model

This standard contains recommendations specific to Web interfaces. "Research-Based Web Design & Usability Guideline" (RBWDUG) is also not compatible with mobile Web interface. But these guidelines have some recommendations, which can be applied on mobile websites. First part of our study is to investigate and expose these recommendations. It is clear that the design of mobile Web interfaces or smart devices could require additional guidance [13] .

Second part of our study is to do a literature search on guidelines, which are tailored towards mobile websites. Nielsen Norman Group's Mobile Usability Guidelines make suggestions for Mobile Web Interfaces. "Usability Of Mobile Websites" (UMW) guidance is based on methodical observations, interviews, user diaries, as well as in-house expert reviews [14]. Their first mobile usability report is announced in 2011 and includes 85 design guidelines for Improving Access to Web-Based Content and Services Through Mobile Devices. In the following year, there have been huge technological improvements in mobile therefore requirements for mobile devices have changed. In 2012, second edition of mobile usability is announced and the number of design guidelines increased to 210 [2].

The World Wide Web Consortium (W3C) announced "Mobile Web Best Practices 1.0" guidance. This guidance outlines 60 guidelines in ten groups for designers and developers to design and deliver content that works well on mobile devices. Its main objective is to improve the user experience of the Web when accessed from mobile devices [15].

Apple's IOS Human Interface Guidelines and Android Human Interface Guidelines are style guidelines for designers who make user interfaces for applications. They tell designers how to design user friendly applications depending on operating systems. These guidelines are not

 ISBN: 978-1-61208-280-6

independent from the platforms. Main objective of these guidelines is to provide consistent design on dependent operating system. They commonly do not make recommendations for mobile websites, which run on different kind of operating systems [16, 17].

B. Guided-Based Usability Evaluation Method

Literature search on guidelines is followed by the categorization step. In the categorization step, guidelines are grouped according to their focuses. List of rule is created in the following categories: accessing, content, homepage, typing, links, searching, navigation, forms, menus, logging, errors, listing and scrolling, images-videos, information.

TABLE I. DEFINITION OF SEARCHING RULE LIST (SEARCH FUNCTION)

RULE	SOURCE	GUIDELINE	GROUP
Rule 1.1	ISO 9241-151 8.5.2.1	Providing a search function	Search Function
	ISO 9241-151 8.5.2.2	Providing appropriate search functions.	
	UMW First Edition Guideline 45.	For smartphones and touch phones with relatively large screens, include a search box on your mobile website	
Rule 1.2	ISO 9241-151 8.5.2.3	Providing a simple search function.	Search Function
	RBWDUG 17.6	Structure the search engine to accommodate users who enter a small number of words.	
Rule 1.3	ISO 9241-151 8.5.2.4	Advanced search	Search Function
	RBWDUG 17.2	Design search engines to search the entire site, or clearly communicate which part of the site will be searched.	
Rule 1.4	ISO 9241-151 8.5.2.8	Search field size	Search Function
	UMW First Edition Guideline 49.	The length of the search box should be at least the size of the average search string. We recommend going for the largest possible size that will fit on the screen (30 characters).	
Rule 1.5	ISO 9241-151 8.5.2.10	Error-tolerant search	Search Function
	UMW First Edition Guideline 50	Preserve search strings between searches. Use auto completion and suggestions	Search Function
	RBWDUG 17.3	Treat user-entered upper and lowercase letters as equivalent when entered as search terms.	
	RBWDUG 17.5	Construct a Web site's search engine to respond to users' terminology.	
Rule 1.6	UMW First Edition Guideline 51.	Do not use several search boxes with different functionalities on the same page	Search Function

In this categorization, guidelines are examined according to mobile limitations, mobile applicability and necessity to be placed on mobile user interface.

There is an example of categorization of guidelines for searching on mobile websites as shown in Table I, Table II, Table III.

TABLE II. DEFINITION OF SEARCHING RULE LIST (SEARCH RESULTS)

RULES	SOURCE	GUIDELINE	GROUP
Rule 2.1	ISO 9241-151 8.5.3.1	Ordering of search results	Search Results
	RBWDUG 17.1	Ensure that the results of user searches provide the precise information being sought, and in a format that matches users' expectations.	
Rule 2.2	ISO 9241-151 8.5.3.2	Relevance-based ranking of search results	Search Results
Rule 2.3	ISO 9241-151 8.5.3.4	Sorting or filtering search results	Search Results

TABLE III. DEFINITION OF SEARCHING RULE LIST (REPEATING AND REFINING SEARCHES)

RULES	SOURCE	GUIDELINE	GROUP
Rule 3.1	ISO 9241-151 8.5.5.1	Giving advice for unsuccessful searches	Repeating And Refining Searches
	UMW First Edition Guideline 52	If the search returns zero results, offer some alternative searches or a link to the search results on the full page	

All searching guidelines, which are applicable on mobile websites, are investigated and a list of searching rules is defined. The searching list is broken down into following 3 rule groups as follows: Search function, search results and repeating and refining searches.

C. Usability Evaluation Tool

In this study, we have developed an evaluation tool, which is coined as Usability Evaluator (UE). This tool keeps a database, which is a list of rules, defined during the analysis process of usability guidelines. The proposed tool provides an environment for usability experts to evaluate mobile websites via list of usability rules. Fig. 2 demonstrates the flow chart for how UE is performed.

 ISBN: 978-1-61208-280-6

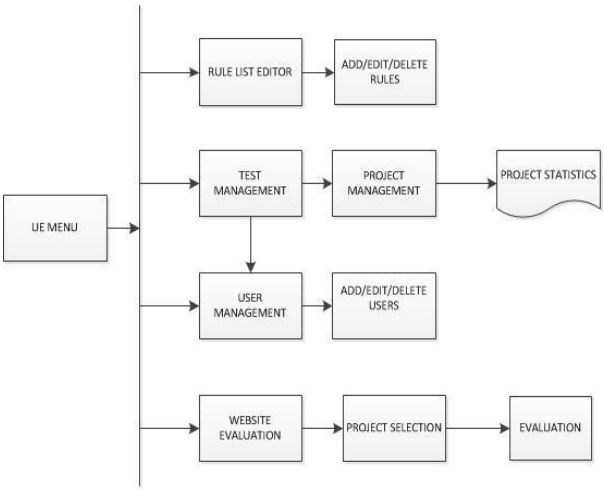


Figure 2. UE Flowchart

UE has "List Editor" menu as shown in Fig.3. Using this menu rules can be prioritized, edited and deleted. Experts can use "Site Evaluation" menu to criticize websites according to rules. UE can calculate percent of success on each rule.

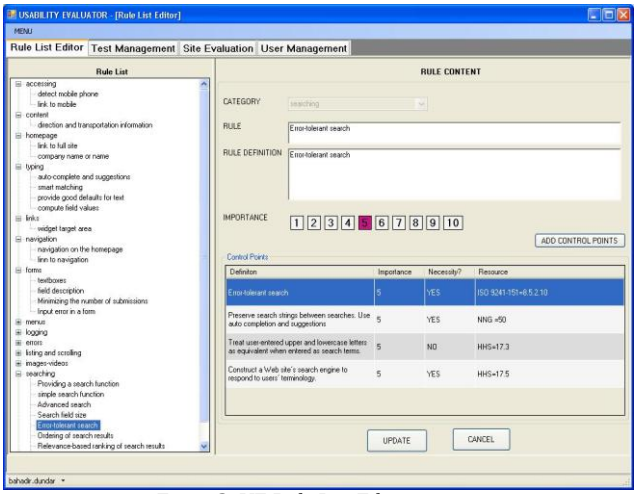


Figure 3. UE Rule List Editor

In the "Test Management" menu, evaluated websites can be monitored. Evaluation statistics can be pulled out depending on the website. EU user authorization can be defined by using "User Management" menu.

IV. GUIDED-BASED USABILITY EVALUATION IN PRACTICE

In this section, we demonstrate practical application of GBUEM on mobile websites. We picked 3 mobile news websites (News X, News Y, News Z) and 3 usability experts from our organization to make the evaluation.

In our study, we focused on touch screen based mobile phones, which have smaller screens than tablets. This evaluation was performed on Samsung Galaxy Note smartphone.

In this practice, weights of the rule categories were determined by the usability experts as shown Table IV. Weighted category scores are obtained by multiplying average category scores and category weights for each category. Total score is the sum of the weighted category scores.

Weighted Category Score (WCS) = Weight of the Category (W) * Category Score (CS) (1)

$$\text{Total Score (TS)} = \Sigma \text{ WCS} \quad (2)$$

TABLE IV. CATEGORY WEGHTS

RULE CATEGORIES	WEIGHTS	News X Category Score	News Y Category Score	News Z Category Score
Access	5 %	50	64	67
Content	10 %	80	72	78
Homepage	5%	50	90	80
Typing	5%	20	40	25
Links	5%	90	93	95
Searching	10%	0	60	0
Navigation	10%	90	93	89
Forms	5%	43	68	40
Menus	5%	100	70	74
Logging	5%	50	90	55
Errors	5%	90	95	88
Listing and Scrolling	10%	83	64	62
Images-Videos	15%	80	55	74
Information	5%	55	65	80
Total Score	100%	67%	71%	64%

UE has the ability to show results as graphical demonstration based on each categories as shown in Fig. 4, Fig. 5 and Fig. 6.

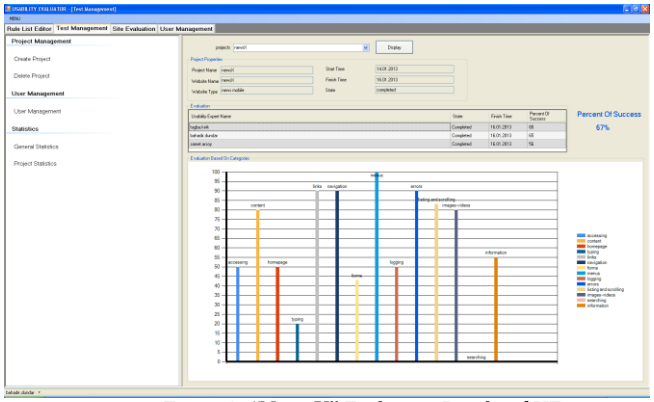


Figure 4. "News X" Evaluation Results of UE

 ISBN: 978-1-61208-280-6

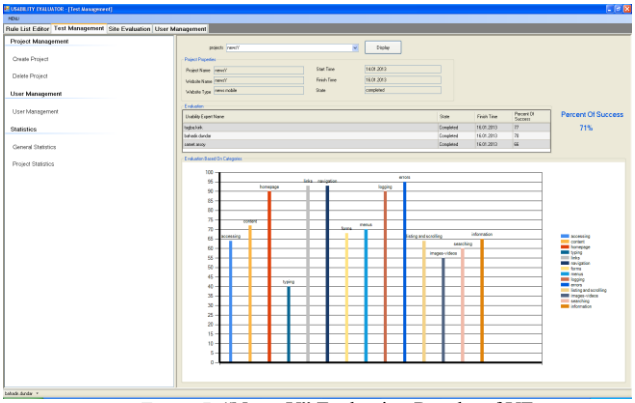


Figure 5. "News Y" Evaluation Results of UE

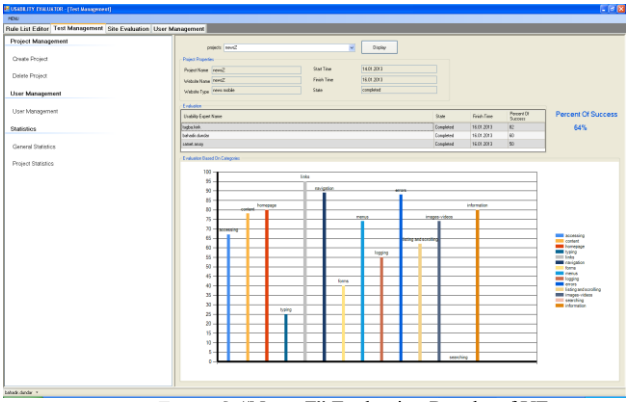


Figure 6. "News Z" Evaluation Results of UE

In this evaluation, usability success rate of these 3 mobile websites are 67%, 71% and 64% eventually. These results are very close to the result of Nielsen Norman Group participant test, which was 64% [14].

The basic findings in this evaluation are as following:

- One of these websites does not detect that user is connected through on a mobile phone and does not direct the user to mobile site.
- 2 of these websites do not include a link to mobile sites.
- One of these websites uses "wap" word in the site title instead of "mobile" or "m" word.
- One of these websites does not include a link to full sites.
- 2 of these websites do not include communication information, about link and privacy policy on their mobile websites.
- 2 of the websites do not have search option. Users are not able to search news in archive and in other sections.
- One of the websites plays the video on the full site if user clicks on the mobile page.
- One of the websites does not include at a link to navigation on every page of mobile website.
- 2 of the websites do not split the list into multiple pages and do not show one page at a time.

V. CONCLUSION AND FUTURE WORK

This study summarized background for mobile usability and gave statistics about mobile and usability issues. In this paper, we have proposed a methodology for user experts in evaluating the usability of mobile websites. We focused on performing this evaluation in a practical and simple way. We discussed the guidelines in the context of mobile usability. We introduced UE and described its operation. We used our tool for evaluating the usability of some mobile websites. Based on the results of this evaluation, we came up with some basic problems. Usability success rate of these websites will be increased by fixing these basic usability problems.

In the process of analyzing guidelines we surveyed literature on mobile usability and determined that there is no standard addressing mobile user interfaces. Some of the guidelines are inadequate and subjective. Considering increasing usage share and importance of mobile websites, there should be a standard for mobile user interfaces. If such a mobile usability standard exists, designers and developers could use this standard for designing consistent mobile user interfaces, which would result in higher mobile user satisfaction. This survey shows us that a standardization study on mobile user interfaces is needed.

This study is intended to help usability professionals in evaluation of mobile environments. We proposed a model to evaluate mobile websites, which didn't involve mobile applications. In our future studies, we will apply this evaluation model on mobile applications by changing the list of rules for UE. Apple's App Store and Google's Android Market user reviews will also be considered to evaluate mobile applications. This model will be applied on not only news websites but also shopping, banking and video sharing websites.

ACKNOWLEDGMENT

We would like to thank Software Testing and Quality Evaluation Center (YTKDM) of TUBITAK, for supporting us and allowing us to use their computer facilities for this study. As always we are really grateful for the help of the extended team of our department.

REFERENCES

[1] Quarterly Mobile Traffic Report, January 8, 2013, [Online]. Available: http://www.walkersands.com /quarterlymobiletraffic. [Accessed: April 16, 2013]

[2] R. Budiu and J. Nielsen, "Mobile Website and Application Usability Second Edition", Nielsen Norman Group, 2012.

[3] A. Wessels, M.Purvis, and S.Rahman "Usability of Web Interfaces on Mobile Devices", IEEE Press, ISBN: 978-0-7695-4367/3, 2011, pp.1066-1067.

[4] ISO/IEC 9241 Ergonomics requirements for office with visual display terminals (VDTs), International Organization for Standardization, Geneva, Switzerland.

[5] J. Nielsen, "Usability 101: Introduction to Usability", Alertbox, January 4, 2012, [Online]. Available: http://www.nngroup.com/articles/usability-101-introduction-to-usability/. [Accessed: April 16, 2013]

[6] J. Nielsen, "How Long Do Users Stay on Web Pages?", Alertbox: September 12, 2011, [Online]. Available:

 ISBN: 978-1-61208-280-6

http://www.nngroup.com/articles/how-long-do-users-stay-on-web-pages/. [Accessed: April 16, 2013]

[7] R. Inostroza, C. Rusu, S. Roncagliolo, C. Jiménez, and V. Rusu,"Usability Heuristics for Touchscreen-based Mobile Devices", IEEE Press, ISBN: 978-0-7695-4654-4/12, 2012, pp. 662-667.

[8] J. Nielsen, "Mobile Content is Twice as Difficult," Alertbox, February 28, 2011, [Online]. Available: http://www.nngroup.com/articles/mobile-content-is-twice-as-difficult/. [Accessed: April 16, 2013]

[9] M. Rauch, "Mobile Documentation: Usability Guidelines, and Considerations for Providing Documentation on Kindle, Tablets, and Smartphones", IEEE Press, ISBN: 978-1-61284-779-5/11, 2011, pp. 1-13.

[10] A. S. Tsiaousis and G. M. Giaglis, "Evaluating the Effects of the Environmental Context-of-Use", IEEE Press, ISBN: 978-0-7695-3260-8/8, 2008 , pp. 314-322

[11] H. Liang, H. Song, Y. Fu, X. Cai and, Z. Zhang, "A Remote Usability Testing Platform for Mobile Phones" IEEE Press, ISBN: 978-1-4244-8728-8/11, 2011 , pp. 312-316

[12] F. Nayebi, J. M. Desharnais, and A. Abran, "The State Of The Art Of Mobile Application Usability Evaluation", IEEE Press, ISBN: 978-1-4673-1433-6/12, 2012, pp. 1-4

[13] ISO 9241 Ergonomics of human-system interaction - Guidance on World Wide Web user interfaces International Organization for Standardization, Geneva, Switzerland.

[14] R. Budiu and J. Nielsen, "Usability of Mobile Websites: 85 Design Guidelines for improving Access to Web-Based Content and Services through Mobile Devices," Nielsen Norman Group, 2011.

[15] J. Rabin and C. McCathieNevile, "Mobile Web Best Practices 1.0" W3C. July 29, 2008, [Online]. Available: http://www.w3.org/TR/mobile-bp/. [Accessed: April 16, 2013]

[16] Apple Inc., iOS Human Interface Guidelines. [Online]. Available: http://developer.apple.com/library/ios/documentation/userexperience/conceptual/mobilehig/MobileHIG.pdf. [Accessed: April 16, 2013]

[17] Google, Android Developer's Guide. [Online]. Available: http://developer.android.com/guide/index.html. [Accessed: April 16, 2013]

 ISBN: 978-1-61208-280-6

A New Technology to Adapt The Navigation

Rim Zghal Rebaî, Corinne Amel Zayani and Ikram Amous
MIRACL
ISIMS, El Ons City, Sfax University, Tunis Road Km 10 Sfax-Tunisia
rim_zghal@yahoo.fr, zayani@irit.fr, ikram.amous@isecs.rnu.tn

***Abstract*— In hypermedia systems, adaptive navigation support becomes a necessity because it helps the user to distinguish between relevant and irrelevant links. It can reduce the loss of time and resolves the disorientation problem. For this, several adaptive navigation technologies are proposed to be applied on simple links to support the user along his navigation. In this paper, we propose an adaptive navigation method based on a new adaptive navigation technology called "Extended Link Technology", which allows restricting the navigation space by using the XLINK extended links. This technology can be applied on both simple and extended link by taking into account several parameters related to the user and the visited documents.**

Keywords— ***Navigation adaptation; adaptive navigation technologies; XLINK extended links.***

I. INTRODUCTION

Currently, data sources in digital format are growing. Therefore, the volume of data and the number of links that connect these data increase. So, the user may be lost in the huge amount of information and links which makes the access to the relevant information more difficult. Thus, navigation adaptation is the solution that supports the user to find pertinent links to the suitable information. Thereby, the user will not be lost in the hyperspace in front of the large number of links and the disorientation problem can be resolved.

Several adaptive navigation methods and navigation technologies have been proposed to help the user by guiding him from a document to another, providing him with a set of pertinent links leading to the suitable information. The navigation adaptation result varies from one user to another, generally according to his profile (needs, preferences, necessities, etc.). But these methods as proposed in [15][7] and technologies as proposed in [4][5] are applied only on simple links and do not take into account the extended links. So, to better reduce the disorientation problem, the navigation space and the number of pertinent links displayed to the user, we propose an adaptive navigation method based on a new adaptive navigation technology which takes into account the extended links. This method is an extension of our proposed method in [19]. Our proposed technology allows to reduce the number of pertinent links in document by using the XLINK extended links (W3C [17]). So, we will apply on documents both our new technology and the already existing technologies by taking into account several adaptation parameters. These parameters are related not only to the user but also to the visited documents.

In order to realize our method, we propose an algorithm that takes as input the adaptation parameters to outputting the expected navigation adaptation. This algorithm uses two functions: the first function assigns scores to links in order to differentiate them (irrelevant, relevant, less relevant, more relevant, etc.) and the second function applies on document our new adaptive navigation technology in order to extract extended links from the simple links.

To evaluate our method, we perform a series of experiments on the INEX 2007 collection. The results of this evaluation are satisfying and prove the efficiency of our method.

This paper is organized as follows. Section II presents a state of the art of some works dealing with the navigation adaptation. In Section III, we expose an overview of our method. In Section IV, we detail the main algorithm and functions on which our method is based. In Section V, we evaluate our proposed method. Finally, we present the conclusion and our perspectives.

II. STATE OF THE ART

Navigation adaptation is mainly based on the adaptation of documents' links by using the adaptive navigation support technologies [4]. The most known and used technologies are the link hiding technology, the link annotation technology, the direct guidance technology, the link ordering technology and the link generation technology.

The AHA! [6], for example, applies the hiding technology to irrelevant links and the link annotation technology to the remaining links by using different colors depending on the user's model (preferences, knowledge). Hypadapter [9] introduced the link ordering technology; the idea is to put links in order of relevance according to the user's model. The direct guidance technology is used to provide only one direct link to the next document to be visited. Among the works that use this technology, we cite Web Watcher [1], ELM-ART [8] and Chiou et al. [7]. The most used technology is the link generation technology. It provides links to the best documents. These latter are identified by means of different methods that vary from one system to another. The adaptive system proposed by Verma et al. [14] calculates and ranks the weight of each web page in the priority of descending order according to click-count, hyperlink weight and most frequent visits to the webpage. Then, it proposes link to the first page. The system proposed in [15] analyzes navigation paths of website visitors to identify the frequent surfing paths and provides the user with a set of links that leads to the next visited web pages. Seo et al. [13] propose two methods based on the generation technology. The first one suggests the next link to be followed by the user and the second one generates quick links as additional entry points into Websites. The system proposed in [20] extracts the

 ISBN: 978-1-61208-280-6

links and the key words from the already visited pages in order to propose a set of links that lead to the relevant pages.

The already mentioned works apply the navigation adaptation only to simple links (lead to only one document or page). As a matter of fact, the number of the pertinent simple links is not reduced. Furthermore, these works adapt the navigation by taking into account only the user's parameters without caring about the documents' parameters. So, we propose a new adaptive navigation technology called "Extended Link Technology" based on the XLINK extended links which reduces the number of the pertinent simple links on document and can be applied on simple and extended links. Our new technology with the already existing technologies is applied to a document before being displayed to the user in order to support him along his navigation.

III. OVERVIEW OF THE PROPOSED METHOD

In order to help the user to easily reach the pertinent information by choosing the best link and to reduce the disorientation problem, we propose a method that adapts the navigation by using: (i) the already existing adaptive navigation technologies [4] and (ii) a new adaptive navigation technology called "Extended link technology". This method is used in our architecture MEDI-ADAPT [18] and it is considered as an extension of our detailed work in [19] which allows identifying the best navigation path between the result documents. So, after identifying the navigation path, we propose to apply to each document, at run time, our new method. This latter consists of 6 steps. The first step is to extract the links of document in order to apply to the irrelevant links the "hiding link technology" (Step 2). The third step is to calculate the scores of links (cf. section IV.2). In the fourth step, we apply the "Extended link technology" (cf. section IV.3). Then in the step fifth, the identified extended links are generated in the document. Finally in the last step, according to the obtained scores in step 3, we apply on simple and extended links the suitable adaptive navigation technologies.

The scores of links, which are calculated in the step 3, have a very important role in our adaptation process. When calculating these scores, we propose to take into account several parameters; parameters related to the user which are extracted from the user profile and parameters related to the links target documents which are extracted from their descriptive meta-documents. In the following, we describe the proposed user profile and meta-documents.

A. The proposed user profile

The user profile is a basic component in adaptive systems. It contains data that describe the user's characteristics. According to Brusilovsky [5]: "user profile is composed of a set of categories: personal data, user's knowledge, interests, history, and preferences". There are different models to represent user profile: the attribute/value model [10], the logical structure based model (usually in a tree form) and the semantic based model via standards e.g. FOAF [3], CC/PP [11], and CSCP [2]. In this paper, as we are interested in navigation adaptation, we consider only the profile's navigation part (the user's navigation history). This latter is built from analysis and updates of previous navigation sessions. We represent the proposed profile as an XML tree as illustrated in the XML schema in Figure 1.

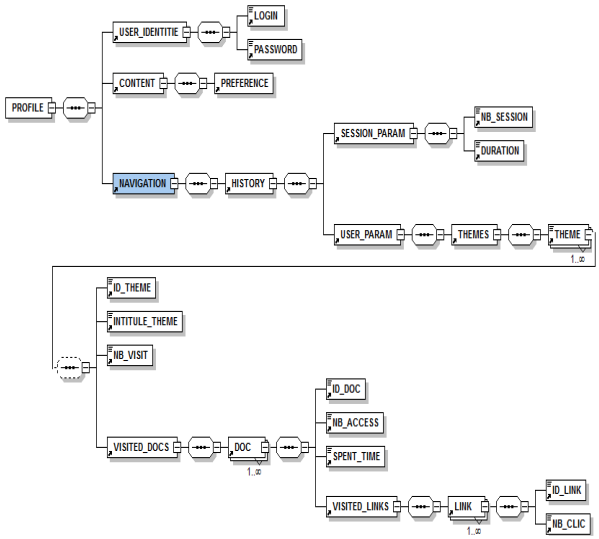


Fig. 1. XML-schema of the proposed user profile

To describe the user's history (HISTORY), we distinguish two parameters: one for the session (SESSION_PARAM) and another for the user (USER_PARAM). The first one consists of the number of sessions (NB_SESSION) and their duration (DURATION). As for the second one, it is made up by the visited documents (VISITED_DOCS) which are identified by (ID_DOC), the number of access to each document (NB_ACCESS), the spent time on each document (SPENT_TIME), the visited links (VISITED_LINKS) and the number of clicks (NB_CLICK) on each link (ID_LINK). In the end of each session, these parameters are updated or stored in the user's profile in order to be taken into account in next sessions.

B. The proposed meta-documents

Meta-documents [1] contain meta-data which are information that can identify a document and describe its content. These information are stored in a document called "meta-document". In this paper, we suggest to describe a document by three parameters stored in the XML meta-document illustrated by the XML schema in Figure 2. These parameters are taken into account in the navigation adaptation process.

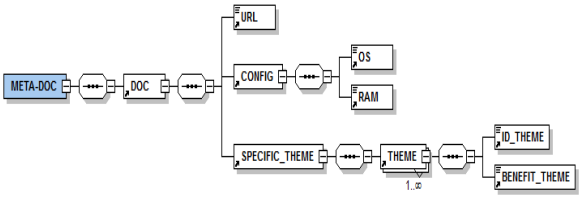


Fig. 2. XML-schema of the proposed meta-document

 ISBN: 978-1-61208-280-6

As we can see in Figure 2, we propose three parameters: (i) the Unified Resource Locator (URL), (ii) the required device configuration (CONFIG) which is limited to the Operating System (OS) and the Random Access Memory (RAM) and (iii) the specific themes (SPECIFIC_THEME), knowing that a document can belong to one or several themes. These themes are consequently used to apply the "Extended link technology": simple links belonging to the same theme are grouped into one extended link. For each theme (ID_THEME), we specify the benefit of the document (BENEFIT_THEME) which depends on the theme. It is similar to the "profit of learning object" proposed in [7]. It is a value ranged from 0 to 1. It can be identified by the document's author based on the relevance of the document's content to each theme.

IV. THE ADAPTIVE XLINK EXTENDED LINK TECHNOLOGY

The originality of our method lies in the use of the new adaptive navigation technology "Extended link technology". This technology is based on the idea of the XLINK extended links. W3C [17] "An extended link is a link that associates an arbitrary number of resources. The participating resources may be any combination of remote and local". This technology allows to reduce the navigation space by reducing the number of links in the document and enables the user to have an idea about the related theme of each link.

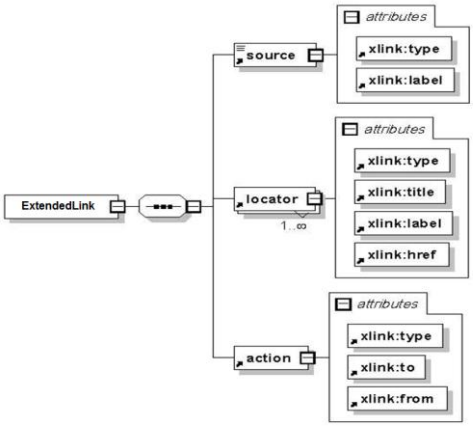


Fig. 3. The XML schema of the generated Extended Links

The basic idea of this technology is to regroup several simple links that belong to the same theme into a single extended link. The founded extended links in document will be generated according to the XML schema illustrated in Figure 3, and take the theme's name as a title. Then, the resources of each extended link are subsequently reordered by using the "ordering link technology" and annotated by the "annotation link technology". We will apply, at run time, this technology on links with the already existing adaptive navigation technologies [3]. In the next section, we will present in detail the main algorithm which performs our proposed method and the two basic functions. The function that calculates the scores of links and the function that applies to document the "Extended link technology".

A. *The main algorithm of the proposed method "ANT"*

To apply our method, we propose the algorithm called "ANT" (Apply the Navigation Technologies) which applies the adaptive navigation technologies on the documents (cf. Table I). This algorithm uses two functions. The first function, called "CLS" (Calculate Link Score) (cf. Table II), which calculates the scores of links. The second function, called "ELT" (Extended Link Technology) (cf. Table III), which allows to apply the Extended link technology.

TABLE I. ALGORITHM ANT

```
1.Algorithm 1: ANT
2. Input: theme,list_themes,selected_doc,
3.          user_history
4. Output:adapted_selected_doc
5. Begin
6.   i=0;
7. for each(link in selected_doc)
8.   begin
9.     if(link not in (theme||list_themes)) then
10.         Hiding_technology(link,selected_doc);
11.      else
12.         begin
13.           i++;
14.           remaining_links[i]=link;
15.       end
16.  end
17.  for each(link in remaining_links)
18.    begin
19.doc_target_link=Determinate_doc_target(link);
20.benefit=Extract_benefit(doc_target_link);
21.Doc_Score=Calculate_Doc_Score(doc_target_link,
         benefit,user_history);
22.Link_Score=Calculate_Link_Score(Doc_Score,
     user_history);
23.  end
24.Extened_link_List=Extended_Link_Technology
       (remaining_links,Link_Score);
25.Generate_Extended_Link(Extended_link_List);
26.end.
```

For each link related to themes other than the requested by the user, the algorithm applies the hiding technology (lines (7) to (10)). Then, for all the remaining links (line (14)), it determines the target document (line (19)), the benefit of this document (line (20)), calculates its score (line (21)) [5], and calculates the link's score (line (22)) by using the function "CLS" (cf. Table II). After that it extracts and generates at run time, if exists, extended links (lines (24) and (25)) by using the function "ELT" (cf. Table III).

 ISBN: 978-1-61208-280-6

B. The CLS function

The CLS function calculates the scores of links. These scores allow to distinguish between links and choose the suitable adaptive navigation technologies. These scores depend on the already cited parameters in the previous sections (cf. Sections III.1 and III.2) and are calculated by the calculation equation (1).

$$Link\_Score(li) = \frac{\sum Doc\_Score(target\_doc)}{nb\_target\_doc} + link\_freqm(li) \quad (1)$$

Doc_Score(target_doc) is the score of the link target document. It is detailed in [19]. In the case of an extended link (having more than one target documents), we proceed to sum the scores of all target documents. *Doc_Score(target_doc))* then divide it by the total number of the target documents *(nb_target_doc). link_freqm(li)* is the average frequency of clicking on the link during all sessions. It is calculated by the calculation equation (2).

TABLE II. CLS FUNCTION

```
1.Function 1: CLS
2. Input: Doc_Score,user_history,link
3. Output:Link_Score
4. begin
5. nb_clic=Extract_nb_click(user_history,link);
6.nb_total_click=Extract_nb_click(user_history);
7.nb_session=Extract_nb_total_session(user_history);
8. calculate link_freqm;          //(cf. Equation2)
9.Return  Link_Score;           //(cf. Equation1)
10.end
```

$$link\_freqm(li) = \frac{\sum \frac{nb\_click(li)}{nb\_total\_click}}{nb\_session} \quad (2)$$

nb_click(li) is the number of clicks on the link in one session, *nb_total_click* is the total number of the visited links and *nb _session* is the total number of sessions.

The CLS function (cf. table II) takes as input the score of the document *Doc_Score*, the history of the user *user_history* and the link link. Firstly, it extracts, from the user's history, the number of clicks on the link and the total number of clicks. Secondly, it calculates, by means of the equation (2) the average frequency of clicking on the link (line (8)). Finally, it uses the equation (1) to calculate the link's score *Link_Score* (line (9)).

C. The ELT function

The ELT function (cf. Table III) extracts, if exist, extended links from the simple links in order to be generated in the document before being displayed to the user. It takes as input the simple links of a document and their scores *Link_Scores*.

TABLE III. ELT FUNCTION

```
1.Function 2: ELT
2. Input: links,Link_Scores
3. Output:Extended_link_list
4. begin
5. Themes=Extract_themes(links);
6. j=0;
7. i=0;
8. for each(Theme in Themes)
9.   begin
10     j++;
11.       for each(link in links)
12        begin
13          if(link in Theme) then
14.              begin
15.                i++;
16.                Extended_link_list[j][i]=link;
17.            end
18.       end
19.  end
20.Extended_link_list=Reorder_links(Extened_link_l
ist,Link_Scores);
21.Annotate_links(Extended_link_list,Link_Scores);
22.end.
```

The basic idea of this function is to regroup all simple links that belong to the same theme into a single extended link (from line (8) to line (19)). Then, the resources of each extended link are reordered and annotated according to their scores (lines (20) and (21)).

V. EVALUATION

In order to evaluate our proposed method, we use a corpus of 110000 documents of INEX 2007 [16] French version, which is a part of the collection WIKIPEDIA XML. Documents in this corpus contain only XLINK simple links.

Given a user launches this query: "Documents related to the norms and computer standards", the system provides 11 documents called (XML, ASCII, SGML, HTML, XSL, ANSI, 10B5, 10BT, 10B2, Vietnam, SAMP)

When we apply the "Extended link technology" to the XSL document, for example, we will have 2 extended links. A part of the obtained result is shown in table IV (Knowing that the XML and XPATH documents belong to the same theme named "XML").

 ISBN: 978-1-61208-280-6

TABLE IV. EXAMPLE OF AN EXTENDED LINK

Post-adaptation extended link	<ExtendedLink xmlns:xlink="http://www.w3.org/1999/xlink"> <source xlink:type="resource" xlink:label="source">XML</source> <locator xlink:type="locator" xlink:label="destination" xlink:href= "3338.xml" xlink:title="XPATH" /> <locator xlink:type="locator" xlink:label="destination" xlink:href ="3332.xml" xlink:title="XML" /> <action xlink:type="arc" xlink:from="source" xlink:to="destination" />… </ ExtendedLink>

We evaluate the use of our proposed method and especially the "Extended Link Technology". This evaluation is performed by computing the obtained links number firstly without navigation adaptation, secondly with well-known (The already existing) navigation adaptation technologies and thirdly with these latter and our proposed technology "Extended Link Technology" (cf. Table V).

TABLE V. NUMBER OF LINKS IN THE RESULT DOCUMENTS

Document Title	without adaptation	with adaptive navigation technologies	with the extended link technology
XML	39	27	6
ASCII	18	2	1
SGML	10	3	3
HTML	72	10	5
XSL	5	5	2
ANSI	4	1	1
10B5	10	3	2
10BT	9	3	2
10B2	13	4	2
Vietn.	5	1	1
SAMP	10	2	1

The variation in the number of links is illustrated in Figure 4. The obtained results show that the use of the well-known navigation adaptation technologies reduces the number of links from 195 to 61. But by applying our proposed technology "Extended Link Technology" we achieve the low number of pertinent links (from 195 links to 26). This means that, by applying our technology with the already existing technology the number of links can reduce to 86.666667%.

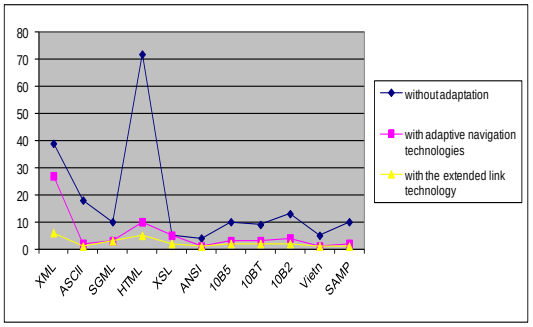


Fig. 4.The variation in the number of links

To evaluate the user's satisfaction, we have given him two document versions: a version without applying our technology and a version with applying our technology. Then, he tries to indicate the relevant links and the irrelevant ones. For him, these latter are links that should have remained simple and not hidden within an extended link. Table VI illustrates the number of relevant links and the number of the irrelevant links.

TABLE VI. NUMBER OF RELEVANT AND IRRELEVANT LINKS

Document title	Number of links without adaptation	Number of links with our technology	Number of relevant links	Number of irrelevant links
XML	39	6	30	3
ASCII	18	1	17	0
SGML	10	3	5	2
HTML	72	5	60	7
XSL	5	2	3	0
ANSI	4	1	3	0
10B5	10	2	6	2
10BT	9	2	6	1
10B2	13	2	8	3
Vietn.	5	1	4	0
SAMP	10	1	8	1

Based on the number of links shown in Table VI, we calculated the precision of the extended links in each document (cf. Figure 5).

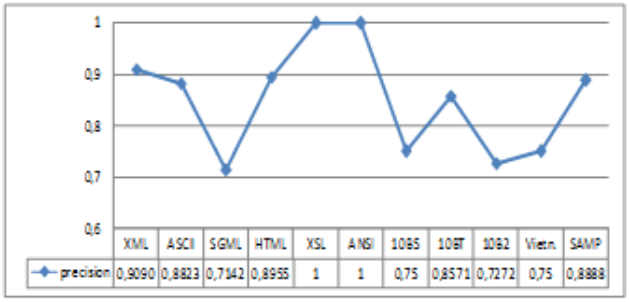


Fig. 5. The precision rates of the extended links

The evaluation of the user's satisfaction allowed us to obtain 0.85223 as an average precision. This can confirm the user's satisfaction. So, the user will have a limited number of well annotated links and can have from the extended links title an idea about the themes of the target documents. That's why; the probability of being disoriented is very limited.

VI. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a method of navigation adaptation which is based on the already existing adaptive navigation technologies and on a new adaptive navigation technology called "Extended Links Technology". This new technology is mainly based on the idea of the XLINK extended

 ISBN: 978-1-61208-280-6

links and can be applied at run time to simple and extended links.

As it is shown in the evaluation, our new adaptive navigation technology allows to reduce the number of pertinent links in document. Thus, the navigation space and the disorientation probability can be reduced.

In the continuation of our work, we aim firstly to evaluate our method with more than one user and evaluate their satisfaction. Secondly, we intend to improve our proposed method by taking into account other parameters. Finally, we are going to suggest and implement a learning method that reduces the profile to the most relevant content.

REFERENCES

[1] I. Amous, A. Jedidi, and F. Sedes, “A Contribution to Multimedia Document Modeling and Querying”, Multimedia Tools and Applications, Vol. 25, n°3, 2005, pp. 391-404

[2] S. Buchholz, T. Hamann, and G. Hübsch, “Comprehensive Structured Context Profiles (CSCP): Design and Experiences”, In PERCOMW '04 Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004, pp. 43-47

[3] D. Brickley and L. Miller, “Foaf vocabulary specification”, Technical report, FOAF project 2007

[4] P. Brusilovsky, “Methods and techniques of adaptive hypermedia”, In User Modeling and User-Adapted Interaction, 1996, pp. 87-129

[5] P. Brusilovsky, “Adaptive hypermedia”, In User Modeling and User Adapted Interaction, Vol.11, 2001, pp. 87-110

[6] P. Brusilovsky et al., “AHA! The Adaptive Hypermedia Architecture”, In Proceedings of the ACM Hypertext Conference, 2003, pp. 81-84

[7] Ch. Chuang-Kai, C.R.T. Judy, Gwo-Jen, H. Shelly Heller, “An adaptive navigation support system for conducting context-aware ubiquitous learning in museums”, In Journal: Computers & Education, Vol. 55, 2, 2010, pp. 834-845

[8] W. Gerhard and P. Brusilovsky, “ELM-ART: An adaptive versatile system for Web based instruction”, International Journal of Artificial Intelligence in Education, Vol . 12, 2001, pp. 351-384

[9] H. Hubertus, B. Heinz-Dieter, G. Rul, “Hypadapter: An adaptive hypertext system for exploratory learning and programming”, In User Modeling & User-Adapted Interaction, Vol. 6, 1996, pp. 131-156

[10] J. Kay, “The um toolkit for cooperating user modeling”, In User Modeling and User-Adapted Interaction, Vol. 4(3), 1995, pp. 149-196

[11] G. Klyn et al., “Composite capability/preference profiles (cc/pp)”, Structure and vocabularies 1.0, Technical report, World Wide Web Consortium (W3C), W3C Recommendation 2003

[12] A. Robert, F. Dayne, J. Thorsten, and M. Tom, “WebWatcher: A learning apprentice for the World Wide Web”, In Proceeding Of AAAI Spring Symposium on Information Gathering from Distributed, Heterogeneous Environments. AAAI Press. 6-12 (1995)

[13] J. Seo, F. Diaz, E. Gabrilovich, V. Josifovski, and B. Pang, “Generalized Link Suggestions via Web Site Clustering”, In WWW '11 Proceedings of the 20th international conference on World wide web, 2011, pp. 77-86

[14] S. Verma, S. Patel, and A. Abhari, “Adaptive web navigation”, In SpringSim '09 Proceedings of the 2009 Spring Simulation Multiconference, paper n° 126

[15] Y-T. Wanga and J.T. Lee. Anthony, “Mining Web navigation patterns with a path traversal graph”, In Expert Systems with Applications, 2011, Vol.38, pp.7112-7122

[16] http://www-connex.lip6.fr/~denoyer/wikipediaXML/, [Retrieved: 4, 2013]

[17] http://www.w3.org/TR/xlink11/, [Retrieved: 4, 2013]

[18] R. Zghal, C. Zayani, and I. Amous, “MEDI-ADAPT: A distributed architecture for personalized access to heterogeneous semi-structured data”, WEBIST, Prtugal, 4. 2012, pp. 259-263

[19] R. Zghal, C. Zayani, and I. Amous, “An adaptive navigation method in semi-structured data”, In ADBIS, Poland, 9. 2012, pp. 207-215

[20] T. Zhu, R. Greiner, and G., Haeubl, “Learning a model of a web user's interests”, In 9th International Conference on User Modeling ,2003, pp. 65-75

 ISBN: 978-1-61208-280-6

Optimal Malicious Agreement in a Virtual Subnet-based Cloud Computing Environment

Kuo-Qin Yan, Hsueh-Hsun Huang
Department of Business Administration
Chaoyang University of Technology
Taiwan, R.O.C.
{kqyan; s9937902}@cyut.edu.tw

Shu-Ching Wang*, Shun-Sheng Wang
Department of Information Management
Chaoyang University of Technology
Taiwan, R.O.C.
{scwang; sswang}@cyut.edu.tw)

***Abstract*—Fault-tolerance is an important research topic in the context of distributed systems. In a distributed system, the cooperative tasks must achieve agreement before performing certain tasks. Nowadays, there are a lot of application services on the cloud computing environment. However, mobile cloud computing is widely accepted as a concept which can significantly improve a user's experience when accessing mobile services. The *Byzantine agreement* (BA) problem is a fundamental problem in fault-tolerance with regard to distributed systems. In previous studies, the BA algorithm is designed using traditional network topology. However, these do not perform well in the context of mobile cloud computing. In order to increase the capability of faulty tolerance and ensure network security, it is necessary to provide a stable mobile cloud service environment. To enhance the reliability of a virtual subnet-based cloud computing environment, a new protocol known as an *optimal malicious agreement* (OMA) is proposed to solve the BA problem in this study. OMA uses the minimum number of message exchanges to make all correct nodes agree on a common value and can tolerate the maximum number of faulty components.**

***Keywords*-**Byzantine Agreement; Fault tolerant; Distributed system; Mobile cloud-computing; Virtual subnet

I. INTRODUCTION

Cloud computing has become a significant technology trend as many applications in the context cloud computing increase convenience for users [10]. Furthermore, the concept of mobile cloud computing inherently provides for the advantages of cloud computing available for users but will provide additional functionality to the cloud as well. Mobile cloud computing will help to overcome limitations of mobile devices, particularly with regard to processing power and data storage [3,5]. However, one of the fundamental mobile cloud computing issues is reliability, where the target mobile nodes connected to the mobile cloud service provider must listen to specific tasks from the server and application recovery is needed.

As mobile cloud computing has become increasingly popular, network topology has trended toward wireless connectivity. Thus, providing enhanced support for mobile cloud computing. In short, this technological trend has greatly encouraged distributed system design and support to mobile nodes. Virtual subnets have attracted significant attention recently because they require less infrastructure, can be deployed quickly, and can automatically adapt to changes in topology. Therefore, virtual subnets suit military communication systems, emergency disaster rescue operations and law enforcement [1]. These, in particular, have brought cloud-computing technology to the mobile cloud computing domain [3,5].

The reliability of the mobile node is one of the most important aspects with regard to the virtual subnet. In order to provide a reliable virtual subnet-based cloud computing environment, a mechanism to allow a set of mobile nodes to agree on a value is required. The *Byzantine agreement* (BA) problem is one of the most fundamental problems [2,9] with regard to reaching an agreement value in a distributed system. The original BA problem defined by Lamport *et al.* [4] is as follows:

(1) There are n nodes in a synchronous distributed system; where n is a constant and $n \geq 4$.
(2) All nodes can communicate with each other through a reliable fully connected network.
(3) One or more of the nodes might fail, so a faulty node may transmit incorrect message(s) to other nodes.
(4) After message exchange, all correct nodes should reach a common agreement, if and only if the number of faulty nodes t is less than one-third of the total number of nodes in the network, or $t \leq (n-1)/3$.

The solutions define a protocol which can reach agreement by using the minimal rounds of message exchange to obtain the maximum number of allowable faulty capability. The problem tackled in this paper involves helping the correct nodes to achieve agreement with underlying n-nodes in a virtual subnet-based cloud computing environment. The source node chooses an initial value to start with and communicates with others by exchanging messages. The nodes have reached an agreement if the scenario satisfies the following conditions [4]:

(Agreement): All correct nodes agree on a common value.
(Validity): If the source node is correct, then all correct nodes shall agree on the initial value which the source node sent.

In previous studies, the BA algorithm was designed for use in a traditional network topology [2]. However, these do not perform well in a virtual subnet-based cloud computing

 ISBN: 978-1-61208-280-6

environment. In order to increase the capability of faulty tolerance and ensure network security, it is necessary to provide a stable mobile cloud service environment. To enhance fault-tolerance, a new protocol known as *optimal malicious agreement* (OMA) in a virtual subnet-based cloud computing environment is proposed to solve the BA problem in this study. OMA uses the minimum number of message exchanges to allow all correct nodes to agree on a common value and can tolerate the maximum number of faulty components.

The rest of this paper is organized as follows. Section 2 discusses the topology of a virtual subnet-based cloud computing environment. Section 3 illustrates the concept of OMA. An example of the execution of the proposed protocol is given in section 4. Section 5 proves the correctness and complexity of our new protocols. Section 6 concludes this paper and offers direction for future research.

II. RELATEED WORKS

Nowadays, the virtual subnet is more practical as it provides the ability for nodes to join the network and leave anytime with no impact on the infrastructure. A group of multiple nodes in a virtual subnet is cooperating to achieve specific objectives; each node communicates with other nodes by broadcasting in the virtual subnet, but also leads to severe problems, such as broadcast storm [1]. Many researchers have proposed cluster schemes where broadcasting is limited and use a virtual subnet to improve upon problems related to broadcast storm and to reduce conflicts. However, recently, virtual subnets have been a more important topic than others [1]. The virtual subnet is composed of several groups through an overlapping network approach [1]. Figure 1 shows a topology of a virtual subnet-based cloud computing environment. There are three situations where the nodes communicate with the underlying topology.

Situation1. Nodes in the same group communicate with each other directly through a virtual backbone.

Situation2. Nodes in different groups exchange messages with each other via a virtual subnet or physical communication media (Internet IP based), e.g., host/agent communication.

Situation3. Host/agent node can communicate with cloud main service via physical communication media in the network topology.

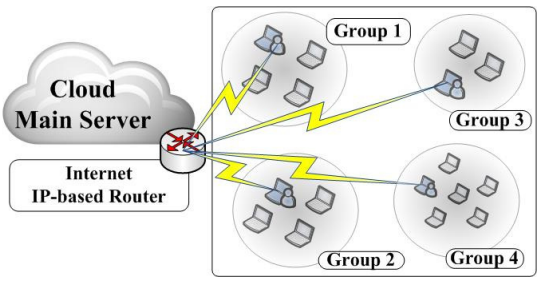


Figure 1. The topology of virtual subnet-based cloud computing environment

In addition, the virtual backbone can be used to: 1) collect topology information for routing; 2) provide a backup route; and 3) multicast or broadcast messages [7]. Hence, a virtual subnet can improve the broadcast storm scenario. In a BA problem, many cases were solved on the assumption of node failure in a fail-safe network [4]. The optimal algorithm for solving the BA problem requires the use of a minimal number of rounds to achieve agreement.

In this study, a new protocol is proposed to solve the BA problem where the communication media in a virtual subnet-based cloud computing environment are reliable but where the node may be faulty through interference from hijackers resulting in the exchanged message exhibiting arbitrary behavior. A protocol reaching agreement in a reliable communication environment involving a traditional fully-connected network was first proposed by Lamport *et al.* [4]. The typical protocol by Fischer [2] can tolerate $f \leq \lfloor (n-1)/3 \rfloor$ faulty nodes in malicious situations and requires σ ($\sigma = f+1$) round(s) to receive enough messages in order to achieve agreement.

However, most of the distributed computing systems may not be fully connected. The network topology has the feature of cluster or group similar to the topology of a virtual subnet. The proposed protocol OMA is used to solve the BA problem underlying a virtual subnet-based cloud computing environment in which the node may fail in a malicious way. When all nodes achieve agreement, the fault tolerance capability has been enhanced even if the communication media has failed between sensor nodes; here, the backbone of the system can be used to provide a backup route [1].

However, the virtual subnet-based cloud computing environment is different than the traditional network, so the previous protocols used in the context of the BA are not suited for the environment this paper proposes. As a result, the new protocol is proposed such that it can be used to solve the BA problem with a malicious fault type in a virtual subnet-based cloud computing environment.

III. THE PROPOSED PROTOCOL

The purpose of the BA protocol is to allow all correct nodes to reach a common agreement in a virtual subnet-based cloud computing environment. For this reason, nodes should exchange messages with all other nodes. Each correct node receives messages from other nodes during a number of rounds of message exchanges. Afterwards, all correct nodes have enough messages to make a decision value, called an agreement value or common value. Then, all correct nodes agree on the same value.

The assumptions, notations and parameters of the proposed protocol OMA are shown as follows:

- Each node in the network can be identified uniquely.
- A node does not know the fault status of other nodes.
- Let n be the total number of nodes in the network.
- Let g be the number of groups in the network and $g \geq 4$.
- Let x be group identifier where $1 \leq x \leq g$ and $g \geq 4$.
- Let n_x be the number of nodes in group Gp_x, $0 \leq x \leq g$. If there are more than $\lceil n_x/2 \rceil$ malicious faulty mobile nodes in Gp_x, then Gp_x will be named the malicious faulty group.

- Let c be the connectivity of the virtual subnet, where c is g–1.
- Let T_{FG} be the total number of malicious faulty groups.
- Let T_{Fn} be the total number of malicious faulty nodes.

In the BA protocol, the first step is to count the number of required rounds of message exchange, which is determined by the total number of nodes at the beginning of protocol execution. Therefore, if the variety of faulty nodes can be determined, then the number of rounds of message exchange can be reduced and then the fault tolerance capability is increased.

The proposed OMA can solve the BA problem due to faulty node(s), which may send incorrect messages to influence the system to reach agreement in a virtual subnet-based cloud computing environment. By using the proposed OMA protocol, all correct nodes in the environment can reach a common agreement which requires θ rounds of message exchange, where $\theta = \lfloor (g-1)/3 \rfloor + 1$.

The proposed OMA protocol is organized in two phases: 1) the message exchange phase and 2) the decision making phase. In the first round of the message exchange phase, the cloud main server sends its initial value to all groups and the receiver node stores the received value in the root of its mg-tree. The mg-tree is a tree structure which is used to store the received message in the message exchange phase from cloud main server [11]. After the first round of the message exchange phase ($\sigma>1$), each node transmits the value at level $\sigma-1$ in its own mg-tree to all other nodes. At the end of each round, the receiver node applies the function RMAJ() to the values received from the same group to obtain a single value. Moreover, each receiver node stores the received messages and the value of function RMAJ() in its mg-tree. RMAJ() is defined in Figure 2.

Subsequently, in the decision making phase, each node outside of the cloud main server reorganizes its mg-tree into a corresponding ic-tree. The ic-tree is a tree structure which is used to store a received message without repeated group names [11]. Therefore, the common value VOTE(s) was obtained by using the function VOTE() on the root s of each mobile node's ic-tree. The detailed steps of the proposed OMA protocol is presented in Figure 2.

IV. AN EXAMPLE OF OMA EXECUTED

An example is given to execute OMA and the virtual subnet-based cloud computing environment is described in Figure 3. There are 22 nodes falling into seven groups. Gp_1 includes P_1 and P_2. Gp_2 includes P_3, P_4, P_5 and P_6. Gp_3 includes P_7, P_8, P_9 and P_{10}. Gp_4 includes P_{11} and P_{12}. Gp_5 includes P_{13} and P_{14}. Gp_6 includes P_{15} and P_{16}. P_{17}, P_{18}, P_{19}, P_{20} and P_{21} belong to Gp_7,:

- The messages are sent from the cloud main server; then, execute OMA.
- The source node C_s (cloud main server) is a malicious faulty server.
- C_s sends 1 to all nodes of Gp_2, Gp_4, Gp_5, Gp_6 and Gp_7 and sends 0 to all nodes of Gp_1 and Gp_3.

In a BA problem with fallible nodes, the worst situation is such that the source node is no longer honest [2]. Put simply, this is the worst case scenario. Suppose the cloud main server is the source node (let it be C_s), which is a malicious fault; this means C_s may arbitrarily send different message values (e.g., replicate command [9]) to different groups. Therefore, in order to solve the BA problem among correct nodes within this example, OMA requires θ ($\lfloor (g-1)/3 \rfloor+1$) rounds of message exchange. pre-execute counts of the number of rounds required before the message exchange phase in OMA. Then, three ($\lfloor (g-1)/3 \rfloor+1 = \lfloor (7-1)/3 \rfloor+1 = 3$) rounds of message exchange are required.

OMA (Source node with initial value v_s)

Definitions:

1. For the virtual subnet, each mobile node has common knowledge of the entire graphic information $\hat{G}$ = (E, Gp), where Gp is the set of groups in the network and E is a set of group pairs (Gp_x, Gp_y) indicating a physical communication medium (the sensing is covered) between group Gp_x and group Gp_y. [7]
2. Each mobile node communicates with all other mobile nodes via the virtual subnet, virtual backbone or physical communication media [1].
3. The node plays sender, receiver or agent, the behavior dictates which kind of transmission is sent[2].
4. The host agent node communicates with the cloud service via a physical communication media (Internet IP based).
5. The host agent node cannot garble the message between the sender node and receiver node; this has been achieved using encryption technology (such as RSA [6]).

Pre-Execute. Computes the number of rounds required $\theta = \lfloor (g-1)/3 \rfloor+1$, where g is the total number of groups in the network.

Message Exchange Phase:

Case $\sigma = 1$, *run*

1. The source node transmits its initial value v_s to each group's nodes.
2. Each receiver node obtains the value and stores it in the root of its mg-tree.

Case $1<\sigma\leq\theta$, *run*

1. Each node without the source node transmits the values at level $\sigma-1$ in its mg-tree to each group's nodes.
2. Each receiver node applies RMAJ on its received messages and stores RMAJ value in the corresponding vertices at level θ of its mg-tree.

Decision Making Phase:

Step 1. Reorganizing the mg-tree into a corresponding ic-tree. (The vertices with repeated group names are deleted).

Step 2. Using function VOTE on the root s of each node's ic-tree, the common value VOTE(s) will obtain.

Function RMAJ(V)

The majority value of the vector $V_i = [v_1, \ldots, v_{n-1}, v_n]$, if the majority exists.

Otherwise, choose a default value ϕ.

Function VOTE(μ)

If the μ is a leaf, then output the value μ.

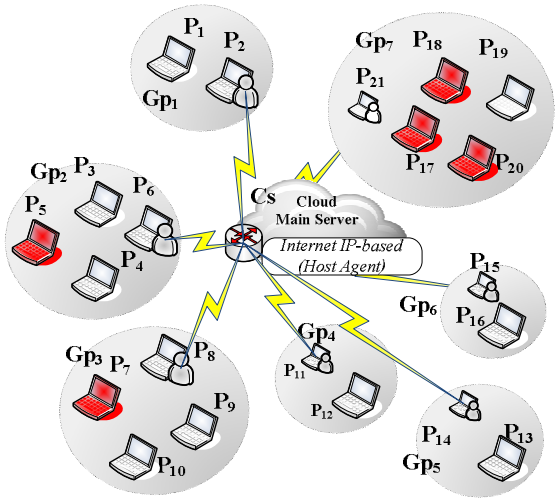
If the majority value does not exist, then output the majority value ϕ.

Otherwise, output the majority m, where $m \in \{0, 1\}$

Figure 2. The proposed OMA protocol

The source node C_s transmits replication commands to all other nodes in the first round of the message exchange phase. The replication command obtained from each correct node is listed in Figure 4. In the σ-th ($1<\sigma\leq\theta$) round of message exchange, except for the Cs, each node transmits RMAJ() values at the (σ–1)-th level in its mg-tree to all other nodes and itself. Subsequently, each receiver node applies RMAJ() to its received messages and stores the received messages and RMAJ() values at the corresponding vertices at level σ of its mg-tree. The mg-tree of the correct node P_1 during the second and final round in the message exchange phase are shown in Figures 5 and 6.

 ISBN: 978-1-61208-280-6



: Correct node : Malicious faulty node

: Host agent node : Cloud main server

Figure 3. The initial status of executing OMA

After the message exchange phase, the tree structure of each correct node is converted from mg-tree to ic-tree by deleting the vertices with duplicated names. The example ic-tree is shown in Figure 7. Eventually, using the function VOTE to root the value *s* for each correct node's ic-tree {VOTE(*s*) = VOTE(*s1*), …, VOTE(*s7*) = 1}, an agreement value 1 can be obtained, as shown in Figure 8. Here, the decision making phase is complete.

	Level 1 *Root s*
Gp_1's correct nodes	0
Gp_2's correct nodes	1
Gp_3's correct nodes	0
Gp_4's correct nodes	1
Gp_5's correct nodes	1
Gp_6's correct nodes	1
Gp_7's correct nodes	1

Figure 4. The mg-tree of each node at first round

Level 1	*Level 2*	*Take*	*RMAJ*
Val(*s*)=1	*s1*	0	(0,0)
	s2	1	(1,1,**0**,1)
	s3	0	(0,**0**,0,0)
	s4	1	(1,1)
	s5	1	(1,1)
	s6	1	(1,1)
	s7	0	(**0**,**0**,1,**0**,1)

Figure 5. The mg-tree of correct P_1 at second round

V. CORRECTNESS AND COMPLEXITY

The following lemmas and theorems are used to prove the correctness and complexity of OMA.

A. *Correctness of OMA*

To prove our protocol's correctness, a vertex is called common [2] if each correct node has the same value for the vertex. That is, if a vertex is common, then the value stored in the vertex of each correct node's mg-tree or ic-tree is identical. When each correct node has a common initial value received from the source node in the root of an ic-tree, an agreement is reached because the root is common. Thus, the constraints, (Agreement) and (Validity), can be rewritten as:

(Agreement'): Root *s* is common and
(Validity'): VOTE(*s*) = v_s for each correct node, if the source node is correct.

To prove that a vertex is common, the term *common frontier* [11] is defined as: when every root-to-leaf path of the mg-tree contains a common vertex, the collection of the common vertices forms a common frontier. Based on these two terms, *common* and *common frontier*, the correctness of OMA can be examined as follows.

Lemma 1: All correct vertices of an ic-tree are common.
Proof: After reorganization, no repeatable vertices are in an ic-tree. At the level T_{FG} +1 or above, the correct vertex α has at least $2T_{FG}$ + 1 children where at least T_{FG} +1 children are correct. The true values of these T_{FG} +1 correct vertices are common, and the majority value of vertex α is common. The correct vertex α is common in the ic-tree, if the level of α is less than T_{FG} +1. As a result, all correct vertices of the ic-tree are common.

Lemma 2: **The common frontier exists in the ic-tree.**
Proof: There are T_{FG} +1 vertices along each root-to-leaf path of an ic-tree in which the root is labeled by the source name and the others are labeled by a sequence of group names. Since, at most, T_{FG} groups can fail, there is at least one correct vertex along each root-to-leaf path of the ic-tree. Following Lemma 1, the correct vertex is common, and the common frontier exists in each correct node's ic-tree.

Lemma 3: Let α be a vertex; α is common if there is a common frontier in the subtree rooted at α.
Proof: If the height of α is 0 and the common frontier (α itself) exists, then α is common. If the height of α is σ, the children of α are all common following the induction hypothesis with the height of the children being σ-1; then, the vertex α is common.

Corollary 1: The root is common if the common frontier exists in the ic-tree.

Theorem 1: The root of a correct node's ic-tree is common.
Proof: By Lemma 1, Lemma 2, Lemma 3 and Corollary 1, the theorem is proved.

Theorem 2: OMA Protocol solves the BA problem in a virtual subnet-based cloud computing environment.
Proof: To prove the theorem, one has to show that OMA meets the constraints (Agreement') and (Validity')
(**Agreement**'): Root *s* is common. By Theorem 1, (Agreement') is satisfied.
(**Validity**'): VOTE(*s*)=*v* for all correct nodes, if the initial value of the source is v_s, say $v = v_s$.

Level 1	*Level 2*	*Level 3*		*Take RMAJ*
s	*s1*	*s11*	0	(0)
0	0(0)	*s12*	0	(0,0,***0***,0)
		s13	0	(0,***1***,0,0)
		s14	0	(0,0)
		s15	0	(0,0)
		s16	0	(0,0)
		s17	1	(***1***,***1***,1,***0***,1)
	s2	*s21*	1	(1)
	1(1,1,1,1)	*s22*	1	(1,1,***1***,1)
		s23	1	(1,***1***,1,1)
		s24	1	(1,1)
		s25	1	(1,1)
		s26	1	(1,1)
		s27	0	(***0***,***0***,1,***0***,1)
	s3	*s31*	0	(0)
	0(0,0,0,0)	*s32*	0	(0,0,***1***,0)
		s33	0	(0,***1***,0,0)
		s34	0	(0,0)
		s35	0	(0,0)
		s36	0	(0,0)
		s37	0	(***0***,***0***,1,***0***,1)
	s4	*s41*	1	(1)
	1(1,1)	*s42*	1	(1,1,***0***,1)
		s43	1	(1,***1***,1,1)
		s44	1	(1,1)
		s45	1	(1,1)
		s46	1	(1,1)
		s47	1	(***1***,***1***,1,***0***,1)
	s5	*s51*	1	(1)
	1(1,1)	*s52*	1	(1,1,***1***,1)
		s53	1	(1,***0***,1,1)
		s54	1	(1,1)
		s55	1	(1,1)
		s56	1	(1,1)
		s57	0	(***0***,***0***,1,***0***,1)
	s6	*s61*	1	(1)
	1(1,1)	*s62*	1	(1,1,***1***,1)
		s63	1	(1,***1***,1,1)
		s64	1	(1,1)
		s65	1	(1,1)
		s66	1	(1,1)
		s67	1	(***1***,***1***,1,***1***,1)
	s7	*s71*	0	(0)
	0(0,0,1,0,1)	*s72*	1	(1,1,***1***,1)
		s73	0	(0,***0***,0,0)
		s74	1	(1,1)
		s75	0	(0,0)
		s76	1	(1,1)
		s77	0	(***0***,***0***,1,***0***,1)

Figure 6. The final mg-tree of node P_1 after the message exchange phase.

Level 1	*Level 2*	*Level 3*		*Take RMAJ*
s	*s1*			
0	0	*s12*	0	(0,0,***0***,0)
	(0)	*s13*	0	(0,***1***,0,0)
		s14	0	(0,0)
		s15	0	(0,0)
		s16	0	(0,0)
		s17	1	(***1***,***1***,1,***0***,1)
	s2	*s21*	1	(1)
	1			
	(1,1,1,1)	*s23*	1	(1,***1***,1,1)
		s24	1	(1,1)
		s25	1	(1,1)
		s26	1	(1,1)
		s27	0	(***0***,***0***,1,***0***,1)
	s3	*s31*	0	(0)
	0(0,0,0,0)	*s32*	0	(0,0,***1***,0)
		s34	0	(0,0)
		s35	0	(0,0)
		s36	0	(0,0)
		s37	0	(***0***,***0***,1,***0***,1)
	s4	*s41*	1	(1)
		s42	1	(1,1,***0***,1)
		s43	1	(1,***1***,1,1)
		s45	1	(1,1)
		s46	1	(1,1)
		s47	1	(***1***,***1***,1,***0***,1)
	s5	*s51*	1	(1)
	1	*s52*	1	(1,1,***1***,1)
	(1,1)	*s53*	1	(1,***0***,1,1)
		s54	1	(1,1)
		s56	1	(1,1)
		s57	0	(***0***,***0***,1,***0***,1)
	s6	*s61*	1	(1)
	1	*s62*	1	(1,1,***1***,1)
	(1,1)	*s63*	1	(1,***1***,1,1)
		s64	1	(1,1)
		s65	1	(1,1)
		s67	1	(***1***,***1***,1,***1***,1)
	s7	*s71*	0	(0)
	0	*s72*	1	(1,1,***1***,1)
	(0,0,1,0,1)	*s73*	0	(0,***0***,0,0)
		s74	1	(1,1)
		s75	0	(0,0)
		s76	1	(1,1)

The tree structure has converted from mg-tree to ic-tree by erasing the vertices with repeated names.

Figure 7. The ic-tree of node P_1.

- VOTE(*s1*) = (0, 0, 0, 0, 0, 1) = 0
- VOTE(*s2*) = (1, 1, 1, 1, 1, 0) = 1
- VOTE(*s3*) = (0, 0, 0, 0, 0, 0) = 0
- VOTE(*s4*) = (1, 1, 1, 1, 1, 1) = 1
- VOTE(*s5*) = (1, 1, 1, 1, 1, 0) = 1
- VOTE(*s6*) = (1, 1, 1, 1, 1, 1)) = 1
- VOTE(*s7*) = (0, 1, 0, 1, 0, 1) = ϕ

VOTE(*s*) = (VOTE(*s1*), VOTE(*s2*), VOTE(*s3*), VOTE(*s4*), VOTE(*s5*), VOTE(*s6*), VOTE(*s7*)) = (0, 1, 0, 1, 1, 1, ϕ) = 1

Figure 8. The common value VOTE(*s*) by correct node P_1.

 ISBN: 978-1-61208-280-6

Since the most of the nodes are correct, they transmit the messages to all others. The value of correct vertices for all the correct nodes' mg-tree is v. When the mg-tree is reorganized to an ic-tree, the correct vertices still exist. As a result, each correct vertex of the ic-tree is common (Lemma 1) and its true value is v. following Theorem 1, this root is common. The computed value VOTE(s) = v is stored in the root for all correct nodes. Thus, (Validity') is satisfied.

B. Complexity of OMA

The complexity of OMA is evaluated in terms of: 1) the minimal number of rounds; and 2) the maximum number of allowable faulty components. Theorems 3 and 4 below will show that the optimal solution was reached.

Theorem 3: OMA requires T_{FG} +1 rounds to solve the BA problem with malicious faults in a virtual subnet-based cloud computing environment where $T_{FG} \leq \lfloor (g\text{-}1)/3 \rfloor$.

Proof: Message passing is required in the *Message Exchange Phase* only. Thus, the message exchange phase is a time consuming phase. Fischer [2] pointed out that t+1 ($t \leq \lfloor (n\text{-}1)/3 \rfloor$) rounds are the minimum number of rounds to get enough messages to achieve BA. The unit of Fischer [2] is nodes, but the unit of the virtual subnet-based cloud computing environment is groups. Here, the number of required rounds of message exchange in the virtual subnet within the cloud computing environment is T_{FG} +1($T_{FG} \leq \lfloor (g\text{-}1)/3 \rfloor$). Thus, OMA requires T_{FG} +1 rounds and this number is the minimum.

Theorem 4: The total number of allowable faulty components by OMA is T_{FG} malicious faulty groups, where $T_{FG} \leq \lfloor (g\text{-}1)/3 \rfloor$.

Proof: The maximal number of allowable faulty nodes to reach BA underlying a fully connected network is f *and* $f \leq \lfloor (n\text{-}1)/3 \rfloor$ [8]. However, the fully connected nature of the virtual subnet-based cloud computing environment is group related; we can suppose a node in Siu *et al.* acts as a group in a virtual subnet-based cloud computing environment [8]. Therefore, $f \leq \lfloor (n\text{-}1)/3 \rfloor$ in Siu *et al.* implies $T_{FG} \leq \lfloor (g\text{-}1)/3 \rfloor$ in a virtual subnet-based cloud computing environment. Therefore, the total number of allowable faulty components by OMA is T_{FG} malicious faulty groups.

As a result, OMA requires a minimal number of rounds and tolerates a maximal number of faulty components to reach a common agreement with correct nodes. Thus, the optimality of the protocol is proven

VI. CONCLUSION

Mobile cloud computing can provide advantages creating better mobile services for users. However, due to the mobility of the network, the nodes of mobile cloud computing may immigrate or emigrate from the network at any time. Furthermore, some of the nodes in the network may be fallible, so the network would not be stable. Notably, the network topology developed in recent years shows a mobile feature [1]. The previous protocols [2,10,11] cannot adapt to solve the BA problem in a virtual subnet of the mobile cloud computing environment. To enhance fault-tolerance, a new OMA protocol is proposed to solve the BA problem herein. OMA uses the minimum number of message exchange rounds to allow all correct nodes to agree on a common value and can tolerate the maximum number of allowable faulty components.

Furthermore, in a generalized case, the fallible components are not only nodes, but also communication media. The OMA protocol may be extended to reach BA in a generalized case underlying the topology of a virtual subnet-based cloud computing environment in the future.

REFERENCES

[1] T.C. Chiang, H.M. Tsai, and Y.M. Huang, "A partition network model for ad hoc networks," Proc. IEEE International Conf. on Wireless and Mobile Computing, Networking and Communications, vol. 3, 2005, pp. 467-472.

[2] M. Fischer and N. Lynch, "A lower bound for the assure interactive consistency," Information Processing Letters, vol. 14, no. 4, 1982, pp. 183-186.

[3] A. Klein, C. Mannweiler, J. Schneider, and H.D. Schotten, "Access schemes for mobile cloud computing," Proc. 2010 Eleventh International Conf. on Mobile Data Management, 2010, pp. 387-39.

[4] L. Lamport, R. Shostak, and M. Pease, "The byzantine general problem," ACM Trans. on Programming Language and Systems, vol. 4, no. 3, 1982, pp. 382-401.

[5] Y.T. Larosa, J.L. Chen, D.J. Dengy, and H.C. Chaoz, "Mobile cloud computing service based on heterogeneous wireless and mobile p2p networks," Proc. IEEE 7th International Wireless Communications and Mobile Computing Conference, 2011, pp. 661-665.

[6] B. Lehane and L. Doyle, "Shared RSA key generation in a mobile ad hoc network," Proc. IEEE Conf. of the Military Communications, vol. 2, 2003, pp. 814-819.

[7] M. Min, F. Wang, D.Z. Du, and P.M Pardalos, "A reliable virtual backbone scheme in mobile ad-hoc networks," Proc. IEEE International Conf. on Mobile Ad-hoc and Sensor Systems, 2004, pp. 60-69.

[8] H.S. Siu, Y.H. Chin, and W.P. Yang, "A note on consensus on dual failure modes," IEEE Trans. on Parallel and Distributed Systems, vol. 7, no. 3, 1996, pp. 225-230.

[9] W.T. Tsai, P. Zhong, E. J. Elston, X. Bai, and Y. Chen, "Service replication with mapreduce in clouds," Proc. IEEE International Sym. on Autonomous Decentralized Systems, 2011, pp. 381-388.

[10] S.S. Wang, K.Q. Yan, and S.C. Wang, "Achieving efficient agreement within a dual-failure cloud-computing environment," Expert Systems with Applications, vol. 38, 2011, pp. 906-915.

[11] K.Q. Yan, S.S. Wang, and S.C. Wang "Reaching an agreement under wormhole networks within dual failure component," International Journal of Innovative Computing, Information and Control, vol.6, no.3, 2010, pp. 1151-1164.

 ISBN: 978-1-61208-280-6

The Anatomy Study of Load Balancing in Cloud Computing Environment

Shu-Ching Wang, Ching-Wei Chen
Department of Information Management
Chaoyang University of Technology
Taiwan, R.O.C.
{scwang; s10114901}@cyut.edu.tw

Kuo-Qin Yan, Shun-Sheng Wang
Department of Business Administration
Chaoyang University of Technology
Taiwan, R.O.C.
{kqyan; sswang}@cyut.edu.tw

Abstract—**In recent years, network bandwidth and quality have improved dramatically, in fact, much faster than the enhancement of computer performance. Cloud computing is an Internet-based resource sharing system in which virtualized resources are provided as a service to users over the Internet. Cloud computing refers to a class of systems and applications that employ distributed resources for use in various applications; these computing resources (service nodes) are utilized over a network to facilitate the execution of complicated tasks. However, Cloud computing resources are heterogeneous and dynamic, connecting a broad range of resources. Thus, when selecting nodes for the execution of a task, the dynamic nature of Cloud computing nodes must be considered. To most effectively utilize the available resources, they have to be properly selected according to the requirements of each task. This study proposes a hybrid load balancing policy to maintain the efficient performance and stability of a Cloud computing environment.**

Keywords-Distributed System; Cloud Computing; Scheduling; Load Balancing; Makespan

I. INTRODUCTION

The Internet is a constantly and rapidly developing global network system, and in order to keep pace with its development, network bandwidth must also constantly develop. Cloud computing is one of such developments, allowing for more applications for Internet users [1,6,7]. Cloud computing environments consist of many commodity nodes that can cooperate to perform specific services.

Users are able to access operational capabilities in Cloud computing environments much faster than they could with Internet applications [3]. However, the infrastructure of the Internet is continuously growing and evolving, progressively allowing the provision of ever more Internet application services. In a distributed computing system, components allocated to different places or in separate units are connected so that they may collectively be used to greater advantage [4]. In addition, Cloud computing has greatly encouraged distributed system design and applications to support user-oriented service applications [7]. Furthermore, many Cloud computing applications, such as YouTube, offer greater user convenience [7].

As technology advances, Cloud computing provides better large-scale resource sharing in a broad-field Internet access environment [1,14]. The limitation of space on conventional distributed systems can thus be eliminated in order to achieve cross-platform compatibility, and to fully exploit the significant resources of all available computers [1,14].

Cloud computing over the Internet provides many applications for users, like Facebook, YouTube, etc. Therefore, determining how best to utilize the advantages of Cloud computing and to ensure that each task is assigned the required resources in the shortest possible time is an important task.

Resources are distributed in a Cloud computing environment, and the stability and performance of each resource varies. In other words, Cloud computing environments are dynamic and composed of heterogeneous resources. Thus, resource selection and task distribution are of particular importance. This study proposes a hybrid load balancing policy that selects an effective node set in the static load balancing stage in order to lower the odds of ineffective nodes being selected, and makes use of the dynamic load balancing stage to ensure that tasks and resources are efficiently balanced. When a node status is changed, a new substitute can be located in the shortest time to maintain execution performance.

The remainder of this paper is organized as follows. Section II focuses on related works. The proposed hybrid load balancing policy is described in Section III. Section IV discusses the design of the simulation experiment. Section V provides the experiment results. Finally, conclusions are presented in Section VI.

II. RELATED WORKS

Cloud computing is a form of distributed computing in which massively scalable IT-related capabilities are provided to multiple external customers “as a service” using Internet technologies [14]. Amazon [10] provides many applications through Amazon Web Services (AWS) as a Cloud computing environment, allowing users to rent required infrastructure or application services [11]. With AWS, users can request computing power, storage and other services, and then access suitable IT infrastructure services on demand [11].

Cloud providers have to achieve a large, general-purpose computing infrastructure and virtualization of that infrastructure for different customers and services in order to provide multiple application services. Furthermore, a software package developed by ZEUS allows Cloud providers to easily and cost-effectively offer every customer a dedicated application delivery solution [15]. The network

 ISBN: 978-1-61208-280-6

framework provided by ZEUS can also be used to develop new Cloud computing methods [13-15]. Based on the ZEUS network framework and the properties of Cloud computing structures, this study uses a three-level hierarchical topology.

However, since a multi-level hierarchical network topology can increase the resource cost of data storage [2], Wang et al. proposed a three-level hierarchical framework [8]. In this framework, service nodes in the third level of the framework are used to execute subtasks, service managers in the second level are used to divide the task into logical independent subtasks and a request manager in the first level is used to assign tasks to a suitable service manager.

The system performance of a Cloud computing environment can be managed and enhanced based on comprehensive status information of each node in the system. There are several methods of collecting the relevant information from nodes, including broadcasting, polling and agents.

Agents have been used extensively in recent years [9]. They have inherent navigational autonomy, and can ask to be sent to other nodes. In other words, an agent does not have to be installed on each visited node, and can collect the relevant information from each node participating in a Cloud computing environment, such as CPU utilization, remaining CPU capability, remaining memory, transmission rate, etc. Therefore, when an agent is dispatched, it does not require any control or connection, and travel flow in maintaining system can be reduced [13]. In this study, the agent is used to gather relevant information, and to reduce wasted resources.

This study also includes a system load balancing policy, and a scheduling algorithm for heterogeneous resources [5, 13]. Generally, load-balancing policies for distributed systems can be categorized into static and dynamic policies [5]. Static load balancing uses simple system data, and based on these data, tasks are distributed through mathematic formulas or other adjustment methods [5]. Dynamic load balancing determines how best to assign tasks to each node in the distributed system. When the system is overloaded, the task causing the overloading will be moved to other nodes and processed for dynamic balance. However, this migration of tasks induces extra system overhead [5].

Therefore, task scheduling will affect the load balancing performance of a system. The following are two typical task-scheduling methods:

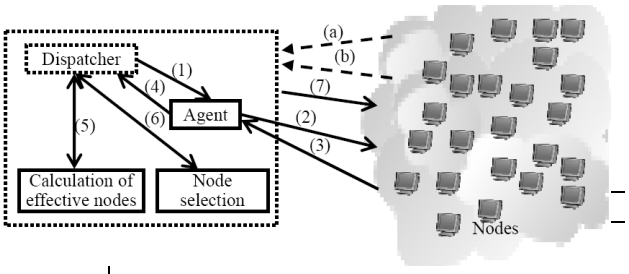
(1) Minimum Completion Time (MCT) assigns each task, together with the minimum expected completion time of each task, to nodes in arbitrary order [5]. This results in some tasks being assigned to nodes that do not have the required minimum execution time for that task [3].

(2) Min-min establishes the minimum completion time for every unscheduled task, and then assigns the tasks to nodes based on the minimum completion offered by each node. The minimum completion time for all tasks is considered, and Min-min can schedule tasks in such a way as to achieve the lowest overall make-span [3].

Many load balancing polices and scheduling algorithms are used to maintain system performance. However, the number of available nodes changes constantly. System performance maintenance, therefore, becomes a complex and difficult process in this dynamic environment. This paper, therefore, proposes a hybrid load balancing policy in order to achieve efficient load balancing.

III. THE HYBRID LOAD BALANCING POLICY

In this section, the proposed hybrid load balancing policy is explained. The structure of the proposed system consists mainly of a dispatcher and nodes. The relationship of roles in this hybrid load balancing policy is described in Figure 1.



(a) When any node can provide its resources, a "join" message and related hardware information will be transmitted to the dispatcher.
(b) When a node can no longer provide resources, it transmits an "exit" message to the dispatcher.

The progressive stages of static load balancing are as follows:
(1) When a request to execute a task is proposed, the dispatcher dispatches an agent to collect the relevant information of each node.
(2) The agent collects the relevant information of the node, such as remaining CPU capability, remaining memory, etc.
(3) The node send its information to the agent.
(4) The agent provides all node-related information to the dispatcher.
(5) The dispatcher builds a table of effective nodes.
(6) The dispatcher selects an effective node set from the table of effective nodes by node selection.
(7) The dispatcher assigns subtasks to the selected nodes.

Figure 1. The interaction of roles

The objectives of the dispatcher include maintaining the load balance, monitoring the status of each node, selecting the nodes for task execution, and assigning tasks for each node. In order to ensure the efficiency of the dispatcher in performing these tasks, an agent will be designed as follows:

(1) The mechanism of an agent mainly collects the relevant information of each node. The information will be provided to the dispatcher to maintain the load balancing of system.

(2) Many factors are considered when a node is selected. Thus, a Value Function (VF) [9] is given to

 ISBN: 978-1-61208-280-6

determine the value of each candidate node, and to provide a reference for selecting effective nodes.

(3) All candidate nodes will be organized into a table of effective nodes. Whenever each node joins or exits the system, the node table will be updated. When any node in the execution aggregate accomplishes its assigned task, it will be transferred to the waiting aggregate, ready for another assignment.

Nodes assist in the execution of tasks in this system. When any node is available or in a busy state, it must transmit its status message to the dispatcher.

In order to maintain system load balancing, this study proposes a hybrid load balancing policy. The proposed policy is carried out in two phases. In the first phase, a static load balancing policy selects an appropriate node for each task. In the second phase, a dynamic load balancing policy, a new node is found to take over the task as soon as the task cannot be completed by the assigned node.

When a request for task execution is made, the task must be divided into several subtasks. The lowest requirements of each subtask determine the threshold on the table of candidate nodes. Nodes passing the threshold are considered as candidate nodes. All candidate nodes can be organized and built into a table of nodes optimized for the proposed task, and the number of required nodes can be determined. If the total number of nodes in the table is smaller than the required number, a portion of subtasks will first be assigned to effective nodes, and the remaining subtasks will be processed when new nodes are added to the table of candidate nodes.

In a dynamic distributed system, the effectiveness of nodes may vary with time. The variation of node status can be identified in two conditions. First, when the dispatcher receives the message that a certain node can no longer provide resources, and second, when the execution of a certain node exceeds the expected time. When either of the above conditions occurs, the dispatcher will launch the agent mechanism for confirmation. If the node remains effective, the distribution of tasks will not be readjusted, but the node's execution of the task will re-estimated. If the node is confirmed to be ineffective, the highest value available node will be selected to replace the ineffective node.

IV. DESIGN SIMULATION EXPERIMENT

In a heterogeneous Cloud computing environment, the performance of nodes varies. In addition, subtasks actually vary in size. Thus, task completion time may vary with the execution order. The properties of both MCT and Min-min are suitable for this experiment, and will be employed and compared with our proposed method. The progression of the experiment is as follows:

(1) The task is divided into 10 independent subtasks.

(2) 10 nodes are selected and assigned tasks by the three different task-scheduling methods.

(3) If any of nodes cannot complete the assigned subtask, new nodes are selected to take over, and the task is then redistributed and re-executed.

According to the above assumptions, this experiment is carried out in two stages. In the first stage, the network simulator, Network Simulation Version 2 (NS-2) [12], is used to dynamically create a Cloud computing environment. In the second, Cloud computing environments with 100, 200, 300, 400, 500, 600, 700, 800, 900 and 1000 nodes are dynamically created with NS-2. Randomly sized data packages are generated and transmitted at a constant bit rate. The transmission rates between the dispatcher and each node are tested. To simulate the heterogeneity of nodes in Cloud computing environments, the CPU capability, memory size, CPU usage and memory usage, and past task completion rate of each node are randomly generated. In addition, the effective time of each node is generated at random, and then multiplied by the past task completion rate in order to reflect the relation between the past task completion rate and the effective time of the node.

According to the Computing Resources (CR) and Amount of Data Transmission (ADT) required to execute the task [7], four scenarios are given:

Scenario 1: CR is large, and ADT is small
Scenario 2: CR is small, and ADT is large
Scenario 3: CR and ADT are large
Scenario 4: CR and ADT are small

In the VF, decision variables can be given different settings, according to the factor focused in the actual application. In the experiment, the available CPU capacity, size of available memory, transmission rate and the past completion rate were the four factors regarded as the threshold for the VF to select nodes and the decision variables for the nodes to estimate their values.

After the decision variables of VF are determined, to make every decision variable comparable, each variable must be quantified. In this experiment, the available CPU capacity and the size of available memory are quantified by the percentage of remaining CPU capacity and memory of each node. Because the transmission rate between the dispatcher and each node is limited to their network bandwidth, the network bandwidth of the dispatcher is taken as the denominator to quantify the transmission rate of each node.

Based on the abovementioned four scenarios, task completion time and number of task redistributions are factors for evaluation. To verify that the nodes selected by VF perform better than those selected by other methods, VFs of different sets of weight are evaluated, and each set is simulated 100 times to obtain objective data. Of the VFs of different weights, the worst weight set that produces the longest completion time (VF-MAX), the average completion time of all weights set (VF-AVG) and the best weight set that produces the minimal completion time (VF-MIN) will be compared to other task-scheduling methods

 ISBN: 978-1-61208-280-6

using the task completion time and number of task redistributions.

V. EXPERIMENT RESULTS AND ANALYSES

The experiment results of Scenario 1 (Figures 2 and 3) indicate that the nodes with the largest CPU resources are selected first, disregarding the execution order for subtasks in MCT. The Min-min selects the combination with the node and subtask that consume the shortest time, so the required task completion time is shorter than in MCT. However, Min-min does not consider the memory provided by the node; subtasks may be re-distributed to nodes with insufficient memory. In addition, the VF considers not only CPU capacity, but also memory and transmission rate and past task completion rate of each node. Even if the task completion time with the worst weight set is close to Min-min, VF still consumes a shorter completion time than Min-min does, and the number of task redistributions are fewer than in Min-min. Thus, VF is more effective in maintaining the load balance.

When the amount of data transmitted is very large, a significant amount of time may be spent on data transmission. This will result in a node being unable to complete the assigned subtask in the effective time, and the subtask will have to be redistributed and re-executed (Scenario 2). As the VF considers these factors, under the various weight sets, fewer task redistributions (Figure 4) and a lower task completion time (Figure 5) are required.

From Figures 6 and 7, it is known that MCT does not consider the execution order. The largest subtask may not be able to find the node with the largest resources, and appropriate nodes cannot be searched for, or assigned to subtasks (Scenario 3). Min-min selects nodes by the shortest completion time instead of the order of subtasks. Therefore, it requires a shorter task completion time than does MCT. As the VF considers multiple factors, nodes that can provide stable resources will be selected first. Even in the worst weight set, task completion time is greater than in Min-min, but with fewer task re-distributions. Therefore, the nodes selected by the VF may not be able to produce the best results in all weight sets; however, they are still effective in maintaining the load balance of a system.

Figures 8 and 9 show that MCT and Min-min only consider CPU capability, and not the memory and transmission rate (Scenario 4). Thus, during task execution, subtasks may be redistributed and a longer task completion time may be incurred. The VF considers multiple factors, so a shorter completion time is required than in almost all the other methods. In addition, task redistribution is almost unnecessary. In other words, the task can be completed in the effective time in almost all cases.

In a Cloud computing environment, the nodes are composed of resources. Since each node has a different hardware structure, nodes cannot be selected based on a single condition (such as available CPU capacity). Therefore, the properties of the task to be executed must be considered. It is known from the above results that when selecting nodes, if the properties of a task and the resources that nodes can provide are not considered, the task to be executed will be repeatedly reassigned and re-executed, thus prolonging the task completion time and lowering the execution performance of system. The VF, however, takes the node resources, transmission rate and past completion rate into consideration. By estimating the value of each node using these factors, the nodes that can provide relatively better resources will be selected. The results of the experiments prove that the hybrid load balancing policy, whether in terms of the best weight set, the worst weight set, or the average time, is far more effective than the other methods in reducing the number of task redistributions and completion time, as well as enhancing the execution performance of the system.

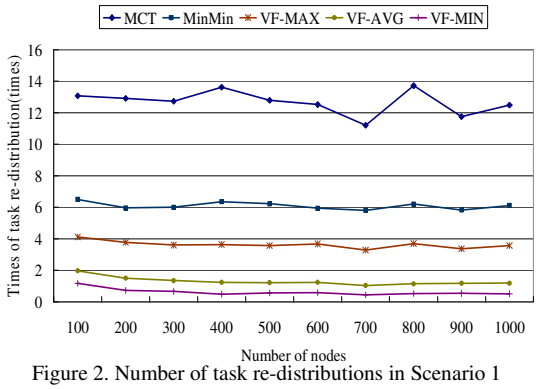


Figure 2. Number of task re-distributions in Scenario 1

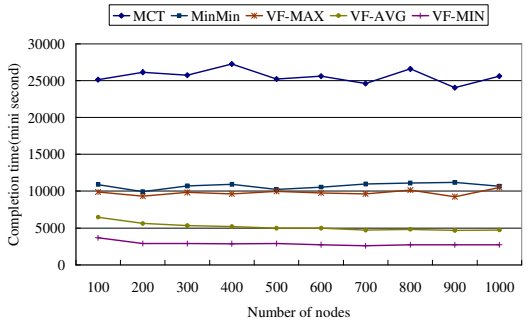


Figure 3. Task completion time in Scenario 1

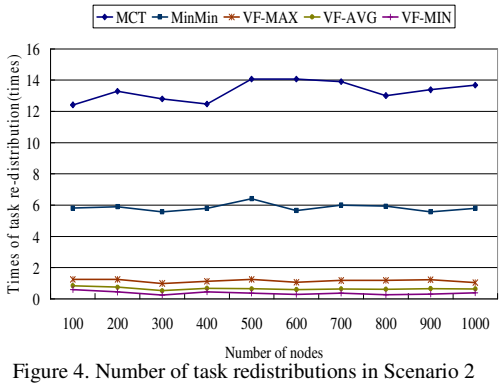


Figure 4. Number of task redistributions in Scenario 2

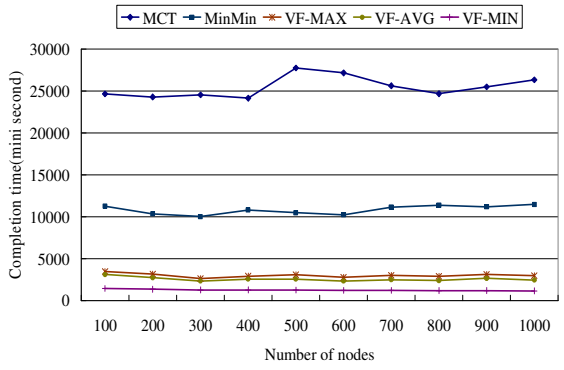


Figure 5. Task completion time in Scenario 2

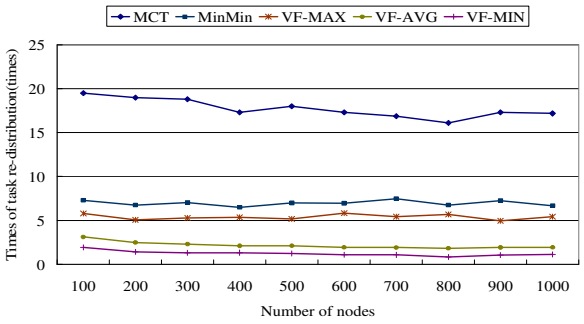


Figure 6. Number of task redistributions in Scenario 3

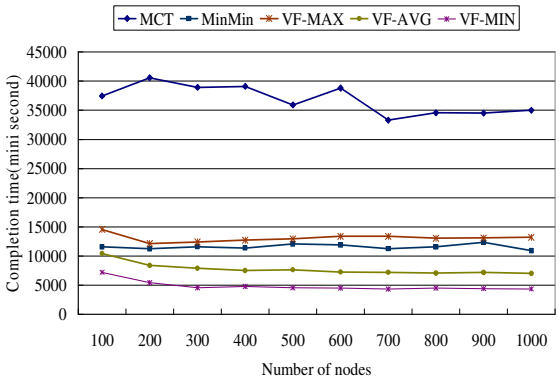


Figure 7. Task completion time in Scenario 3

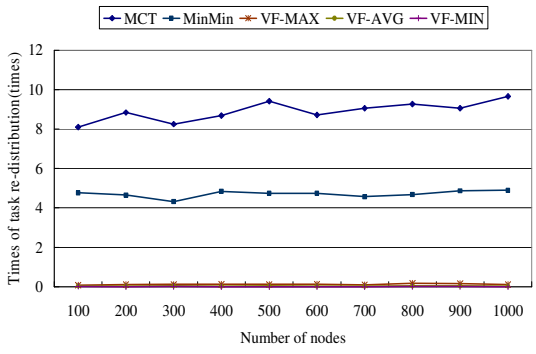


Figure 8. Number of task redistributions in Scenario 4

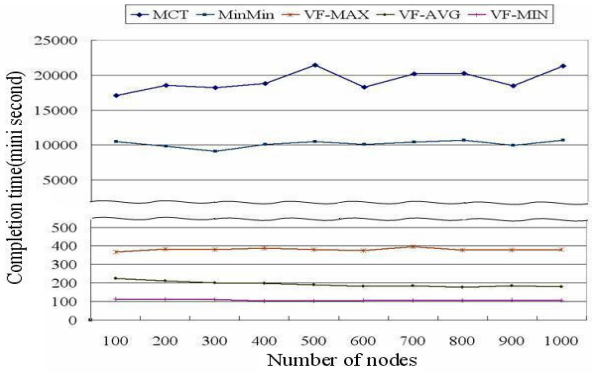


Figure 9. Task completion time in Scenario 4

VI. CONCLUSION AND FUTURE WORK

Cloud computing environments offer many available resources; however, the availability of nodes that provide these resources dynamically changes over time. In this paper, a hybrid load balancing policy is proposed for Cloud computing environments in order to efficiently distribute tasks to available nodes with the required resources for the completion of those tasks in the shortest possible time. The proposed policy is carried out in two phases. In the first phase, a static load balancing policy selects an appropriate node for each task. In the second phase, a dynamic load balancing policy, a new node is found to take over the task as soon as the task cannot be completed by the assigned node.

Since Cloud computing environments are more complicated than traditional distributed systems, it follows that if this policy can achieve efficient load balancing in Cloud computing environments, then it can also solve load balancing issues in other distributed systems.

ACKNOWLEDGMENTS

This work was supported in part by the Taiwan National Science Council under Grants NSC101-2221-E-324-032.

REFERENCES

[1] Y. Gong, Z. Ying, and M. Lin, “A survey of cloud computing,” Lecture Notes in Electrical Engineering, vol. 225, 2013, pp. 79-84.

[2] C.L. Hu and T.H. Kuo, “A hierarchical overlay with cluster-based reputation tree for dynamic peer-to-peer systems,” Journal of Network and Computer Applications, vol. 35, Issue 6, November 2012, pp. 1990-2002.

[3] S. Nesmachnow, F. Luna, and E. Alba, “An efficient stochastic local search for heterogeneous computing scheduling,” Proc. IEEE 26th International Parallel and Distributed Processing Symp. Workshops & PhD Forum (IPDPSW), 21-25 May 2012, pp. 593-600.

[4] N. Olifer and V. Olifer, Computer network: principles, technologies and protocols for network design. John Wiley & Sons, 2006.

[5] B. Sahoo, D. Kumar, and S.K. Jena, “Observing the performance of greedy algorithms for dynamic load balancing in heterogeneous distributed computing system,” Proc. 1st International Conf. on Computing, Communication and Sensor Networks- CCSN,(2012), vol. 62, 2012, PIET, Rourkela, Odisha, pp. 265-269.

 ISBN: 978-1-61208-280-6

[6] A. Vouk, “Cloud computing- issues, research and implementations,” Information Technology Interfaces, June 2008, pp. 31-40.

[7] L.H. Wang, J. Tao, and M. Kunze, “Scientific cloud computing: early definition and experience,” Proc. 10th IEEE International Conf. on High Performance Computing and Communications, 2008, pp. 825-830.

[8] S.C. Wang, W.P. Liao, K.Q. Yan, and S.S. Wang, “Towards a load balancing in a three-level cloud computing network,” Proc. 2010 3rd IEEE International Conference on Computer Science and Information Technology (IEEE ICCSIT 2010), Chengdu, China, 9-11 July 2010, pp. 108-113.

[9] K.Q. Yan, S.C. Wang, C.P. Chang, and J.S. Lin, “A hybrid load balancing policy underlying grid computing environment,” Computer Standards & Interfaces, 2006.

[10] Amazon web services, http://aws.amazon.com/, April 2, 2013.

[11] Application delivery networking, application acceleration, Internet traffic management system: Zeus.com, http://www.zeus.com/, April 2, 2013.

[12] The network simulator - NS-2, http://www.isi.edu/ nsnam/ns/, April 2, 2013.

[13] Load balancing, load balancer, http://www.zeus.com/products/ zxtmlb/index.html, April 2, 2013.

[14] What is cloud computing?, http://www.zeus.com/Cloud_ computing/Cloud.html, April 2, 2013.

[15] ZXTM for cloud hosting providers, http://www.zeus.com/Cloud_ computing/ for_ Cloud providers.html, April 2, 2013.

 ISBN: 978-1-61208-280-6

Cross-Media Retrieval for Music by Analyzing Changes of Mood with Delta Function for Detecting Impressive Behaviours

Yoshiyuki Kato
Faculty of Environment and Information Studies
Keio University
5322 Endo, Fujisawa, Kanagawa, Japan
t10247yk@sfc.keio.ac.jp

Shuichi Kurabayashi
Faculty of Environment and Information Studies
Keio University
5322 Endo, Fujisawa, Kanagawa, Japan
kurabaya@sfc.keio.ac.jp

***Abstract*—This paper proposes a system that retrieves music by accepting a sequence of images as a query representing a change of sentiments in music. The system offers a query model that utilizes image files as a media for describing users' emotional demands for continuous changes of mood in music. This query model offers two types of delta functions, corresponding to music and images. Each delta function measures continuous changes in the corresponding sequential media. Applying the delta functions to the media data generates the values representing changes of moods. Each delta value is normalized distance in the corresponding metric space, thus, comparison of delta values extracted from heterogeneous media data makes it possible to calculate the cross-media relevance score. As a prototype implementation, we have developed a Web-based cross-media retrieval engine that provides an integrated user interface (UI) to create music queries by novices who may submit only rough and simple information.**

Keywords-Cross Media Retrieval; Multimedia Database; User Interface; Mood Analysis;

I. INTRODUCTION

Although music is a very common media in our daily lives, it is difficult to find music satisfying our preferences. This is because music changes in sentiments with time. In order to find the desired type of music, a user must listen to several parts of music in repositories, such as online music stores and personal music players. The main gateways for those online stores are provided as Web-based services, such as Amazon MP3 and Google's Play Music service.

In spite of the fact that young age users tend to select music according to their feelings, users have serious frustration to retrieve their favorite music, due to the lack of web technologies for inputting queries to find continuous media data such as music. Especially for finding recently released music that is unknown to a user, a method to describe a user's ambiguous desires for music is required [1]. Novices need a toolkit that assists its users to form their own queries in a trial-and-error manner. It is desirable to develop an intuitive Web-based toolkit for representing the demands of users for music.

Toward the above objective, this paper proposes a web-based music retrieval method that offers a cross-media query model utilizing "image files" as a media for describing users' emotional demands for continuous changes of mood in music. It is important to develop a stream-oriented query construction method for music, because music changes its content and impression with time. Our query model interprets the perceptional effects of temporal changes in media features such as tonality in music and colors in images. This paper shows a prototype system implementation realizing web-based music retrieval with considering changes of mood.

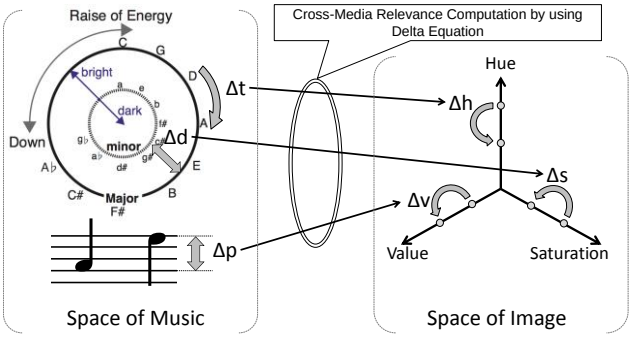


Figure 1. An Conceptual Overview of the Cross-Media Retrieval by using Delta Equation on Each Media Data

Our method achieves the cross-media retrieval by comparing the results of continuous sentiment analyses of music and images. In order to analyze the temporal changes in media data, this system provides two types of delta functions, corresponding to music and images, to generate the sequential values representing changes of mood. Those sequential values represent how the media data changes its sentiment with time. The system calculates the sentiment-oriented relevance score of music and images by comparing the calculated delta values.

Our design principle for this system is to make it possible to search musical contents that are invisible to users by using visible image contents. This concept is highly suitable for the web-based music retrieval because web is a visual media. Thus, a key technology of this system is a cross-media query interpretation method that recognizes how the media changes with time by using several metric spaces to calculate distances between the states of the media and the previous states of the media. For music, we have implemented three metrics based on music tonality analysis methods [2][3]: a tonality metric, a pitch metric, and a major/minor metric. Corresponding to these metrics for music, we have implemented three

 ISBN: 978-1-61208-280-6

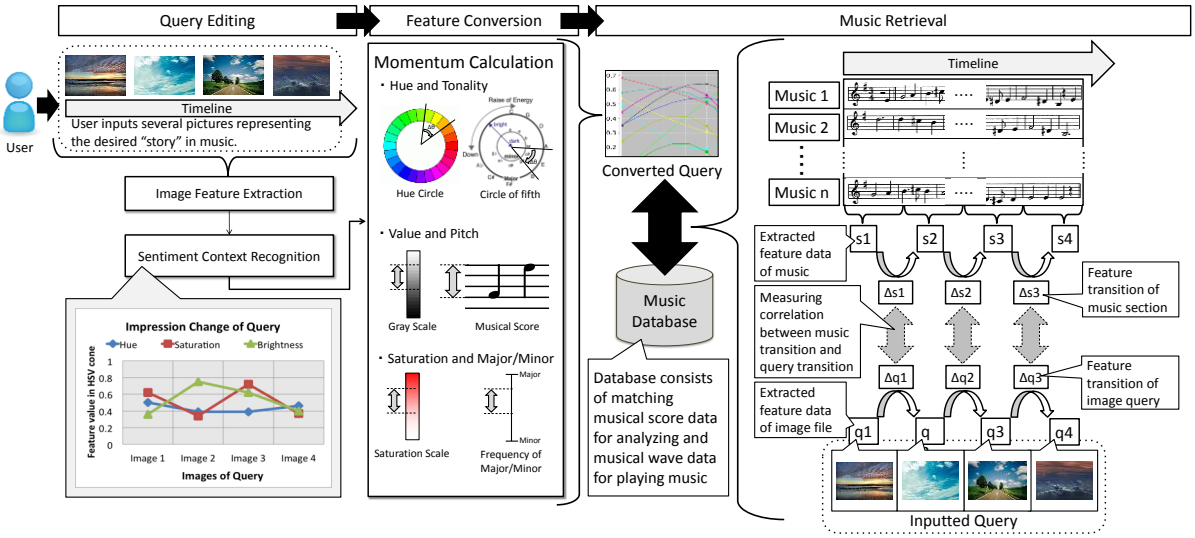


Figure 2. System Architecture for Retrieving Music by Delta Function Analyzing Changes of Mood

metrics for images based on the HSV color space: a hue metric, a saturation metric, and a value metric.

The system transforms invisible changes in the impression of music into visible changes in color and vice versa. Our approach converts "a delta value", which is distance in each space, between two spaces rather than a feature value itself because the system focuses on how the changes of mood affect on human perceptions. The most important feature of the two metric spaces is their configuration as topologically equivalent structures (Figure 1). Each axis in music space has a corresponding axis in the color space. Specifically, tonality is associated with hue, pitch with value, and major/minor with saturation. Thus, a specific distance in music space can be converted into the same distance in color space.

The advantage of this system is an intuitive method for users to edit a query in a trial-and-error manner, depending on their impression. The method makes it possible for users to describe changes in impression of music, which is difficult to represent directly, as a sequence of images with visually enhanced user interface, wherein the order of the images represents a change of impression.

The remainder of this paper is structured as follows. Section II presents motivating example of our query processing. Section III briefly summarizes the related work. Section IV describes the fundamental concept and the system architecture. Section V concludes this paper.

II. MOTIVATING EXAMPLE

In this section, we present a motivating example of our stream-oriented cross-media retrieval. In this example, a user wants to find a music that initially gives a dark impression and gradually makes a brighter impression like a sunrise, but the user does not have clear idea about the title and the artist of the music. In this case, by using our system, the user can retrieve the music with combining four pictures as shown in the left-hand side of Figure 2. The horizontal axis from left to right corresponds to elapsing time. This query represents changes in impression as follows: brightness (value in HSV color model)" is gradually increasing and then decreasing, and hue (type of colors) is stable, and saturation (vividness of colors) changes drastically with time. Hence, according to the relation between a pair of the dual-metrics, the system retrieves music corresponding to changes in impression as follows: pitch is gradually increasing and then decreasing, tonality is stable, and major/minor changes drastically.

III. RELATED WORK

The music information retrieval (MIR) system is a well-known means of helping users to find music by using several intuitive queries [1][4]. A traditional example of the MIR method is a content-based MIR system [5]. As a typical method of the content-based music information retrieval system, query by humming (QBH) makes it possible to use humming as a search key for finding a musical composition [6]. QBH is effective to find music that is familiar to a user, but cannot be applied to find music that is unknown to the user. As described in [7], novel techniques to search for new music that we have not heard are in great demand, because the conventional content-based music retrieval methods do not satisfy such queries.

Ciuha et al. [8] show a music visualization system that utilizes cross-media relationship of colors and tonalities. This system partially supports users in finding new and unknown music. Multi-timescale visualization techniques for displaying the output from key-finding algorithms were presented in [9]. An impression-based music visualization method that utilizes a result of a synesthesia study [10] was proposed in [11]. This uses a color sense of tonality to view the harmonic structure and relationships between key regions in a musical composition.

The most significant difference between the conventional approaches and our approach is that our system focuses on metrics implied by feature transitions in vector space, as the elements representing changes in the impressions of media. We do not use a knowledge base of music and color referring to synesthesia since this

 ISBN: 978-1-61208-280-6

is difficult to personalize for each user by reflecting their individual differences in impressions of media.

IV. System Architecture

In this section, we present our music retrieval system that interprets an image sequence as changes in the sentiment of music. As shown in Figure 2, the system consists of three main components as follows: 1) a query editor, 2) a feature conversion module, and 3) a retrieval engine. The query editor is the front-end module of the system. This module provides a set of operations to prepare and modify image files as a query according to a user's preferences. For example, the system implements an image-editing operator equipped with several color filters to change the overall impression of the image. The core component of the system is a retrieval engine that calculates the correlation between the query and retrieval target items according to their own changes in impression by using the metadata generated in the feature conversion module.

The system models the concept of "change in sentiment" by measuring the distance caused by feature transition of media data. Specifically, the system provides a bridging mechanism between the musical tonality metric space and the HSV color metric space. The bridge converts a distance calculated in the music space into a distance in the image space as keeping its impression factor. For example, in the distance conversion mechanism, hue, which is a type of color, corresponds to tonality, which is a type of music structure. By converting distances between heterogeneous metric spaces, the system realizes cross-media retrieval for stream media such as music.

A. Data Structure

The data structure in this system consists of two data elements, which are the image query and the music. An image-query object Q is defined as follows: $Q := \langle\langle h_1, s_1, v_1\rangle \ldots \langle h_i, s_i, v_i\rangle\rangle$, where h_i is hue data, s_i is saturation data, and v_i is brightness data, in the *i*-th image of the query. The system converts the RGB color values of each image into HSV triples at the pixel level. We define the metric space of images as the HSV color metric space with three axes: hue, saturation, and value. These three elements are significant factors affecting the impression of the image.

Hue represents the differences of color phases such as red, yellow, green, and blue. In the HSV cone of images, hue is represented by angle. The system converts the extracted hue angle in HSV cone of images into a hue scalar *h*, which is a value between 0 and 1. Saturation is vividness of color. In our system, the saturation is an average value of the vividness in an image. The system processes this vividness value of an image into the saturation scalar *s*, which is a value between 0 and 1. Here 0 and 1 represent the lowest value and the highest value, respectively. Value is brightness of color. The prototype system calculates the value (brightness) as an average value of color brightness in an image. Our system processes this brightness value of an image into a value of brightness scalar *v*, which is a value between 0 and 1.

When the proposed system receives a query consisting of several images, the proposed system divides music into sections, and the number of sections is equal to the number of images inputted as the image query. Thus, in this paper, a music object M is defined as follows: $M := \langle\langle t_1, d_1, p_1\rangle \ldots \langle t_i, d_i, p_i\rangle\rangle$, where t_i is tonality data, d_i is deviation data, and p_i is pitch data, in the *i*-th section of the music. We define the metric space of music by three axes: tonality, pitch, and major/minor. These three elements are significant factors affecting to the impression of music.

Tonality is the structure of music, which is composed of sequential musical notes. There are 24 kinds of tonality consisting of 12 major tonalities and 12 minor tonalities. Tonality changes with time in music, and this causes changes in the impression of the music. In music theory, there is the circle of fifths, which defines the distance or similarity between each pair of the 24 tonalities. Each tonality can be represented by an angular value on the circle of fifths. The system processes this angular value into a tonality scalar *t*, which is a value between 0 and 1. Thus, the system converts the distance measured in hue's angle into the distance measured in tonality's angle.

Major/minor refers to a deviation of tonality within a music section. The system calculates the deviation of tonality in a music section, and converts the deviation value into the major/minor scalar *d* that is a value between 0 and 1, which represent the maximum minor deviation and the maximum major deviation, respectively.

Pitch is a value of pitch in a musical score. The system calculates the average of the pitches in a music section, and converts the average value into the pitch scalar *p* that is a value between 0 and 1, which represent the lowest pitch and the highest pitch, respectively.

B. Primitive Functions

For the relevance computation, the proposed system provides a cross-media relevance calculation function and six distance functions. The system calculates how the media data change with time by applying distance functions to each feature extracted from the media data. Then, the system compares the set of distance values to calculate the relevance of the music to the image query. The following three functions are the distance functions for image query:

- The distance in hue between the *i*-th image and the (*i*+1)-th image is $\Delta_{hi} := |h_i - h_{i+1}|$, where *h* is the hue angle in the HSV cone.
- The distance in saturation between the *i*-th image and the (*i*+1)-th image is $\Delta_{si} := |s_i - s_{i+1}|$, where *s* is the saturation coordinate in the HSV cone.
- The distance in value between the *i*-th image and the (*i*+1)-th image is $\Delta_{vi} := |v_i - v_{i+1}|$, where *v* is the value coordinate in the HSV cone.

The following three functions are the distance functions for music:

- The distance in tonality between the *i*-th section and the (*i*+1)-th section is $\Delta_{ti} := |t_i - t_{i+1}|$, where *t* is the tonality angle in the circle of fifths.
- The distance in tonality deviation between *i*-th section and the (*i*+1)-th section is $\Delta_{di} := |d_i - d_{i+1}|$, where *d* is the deviation in tonality.

 ISBN: 978-1-61208-280-6

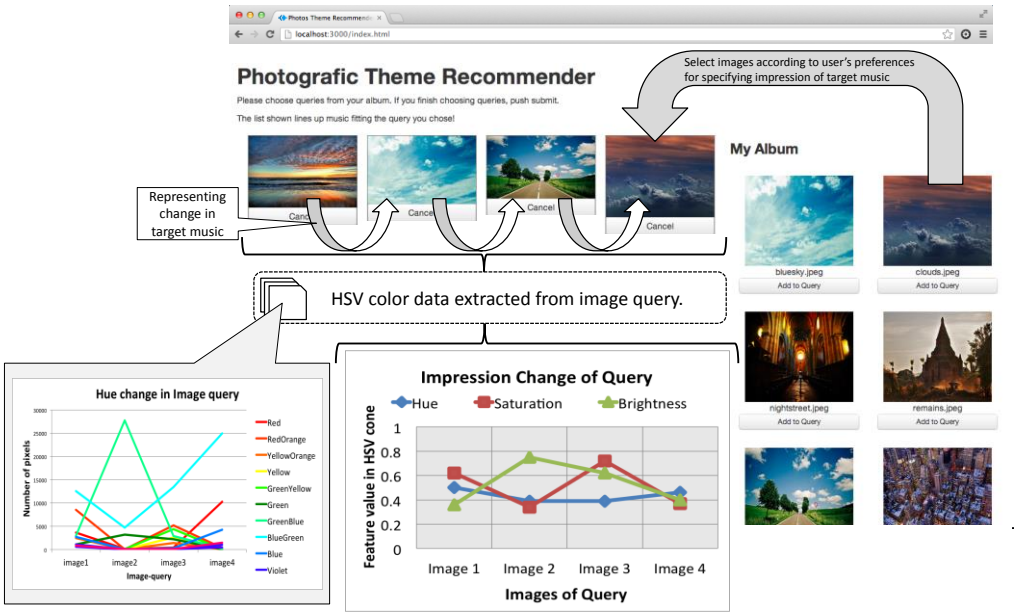


Figure 3. Prototype Implementation that Employs Modern HTML5 Technologies to Edit Image Queries Representing Changes of Mood

- The distance in pitch between the *i*-th section and the (i+1)-th section is $\Delta_{pi} := |p_i - p_{i+1}|$, where *p* is the pitch.

The system provides a fundamental function to calculate the relevance of the music to the query. The function is defined as follows: $(a, b) \rightarrow 1 - |a - b|$, where *a* and *b* form a pair of distance changes according to the dual-metrics relation. The system calculates a correlation value for each pair of metrics by using this function. Moreover, the relevance of the music to the image query is represented as follows:

$$\gamma(\Delta q, \Delta m) := \frac{\sum_{i=1}^{n} \frac{s(\Delta_{hi}, \Delta_{ti}) + s(\Delta_{hi}, \Delta_{ti}) + s(\Delta_{hi}, \Delta_{ti})}{3}}{n} \quad (1)$$

where *n* is the number of images inputted as image-query, as well as the number of divided music sections.

V. PROTOTYPE IMPLEMENTATION

We have implemented a prototype of the proposed system. The prototype system is implemented using HTML5 Canvas, jQuery UI, and Backbone.js as shown in Figure 3. The system implemented consists of three modules: the query editor, the feature conversion module, and the music retrieval engine. The query editor is the main user interface. Users can edit a query by selecting four images and revising the order of the selected images. The feature conversion module generates a query by converting the feature of the inputted image sequence into changes of sentiment in the target music. The music retrieval engine calculates the relevance of candidate music to the converted query.

VI. CONCLUSION REMARKS

We have proposed a cross-media retrieval system for music. The system not only analyzes the changes of mood in candidate music and the sequence of images but also calculates the relevance of the music to the images by using a specially designed metric space for change calculation. As future work, we plan to develop a social-network-based query recommendation by this approach.

REFERENCES

[1] M. Goto and K. Hirata, "Recent studies on music information processing," Acoust. Sci. Technol., vol. 25, no. 6, 2004, pp. 419–425.

[2] D. Temperley, Music and Probability, MIT Press, 2007.

[3] C. L. Krumhansl, Cognitive Foundations of Musical Pitch, Oxford Univ. Press, 1990.

[4] R. Typke, F. Wiering, and R. Veltkamp, "A Survey of Music Information Retrieval Systems," Proc. ISMIR 2005, Univ. of London, 2005, pp. 153–160.

[5] Y. Hijikata, K. Iwahama, and S. Nishida, "Content-based Music Filtering System with Editable User Profile," Proc. ACM SAC'06, ACM Press, 2006, pp. 1050–1057.

[6] A. Ghias, J. Logan, D. Chamberlin, and B. C. Smith, "Query by Humming: Musical Information Retrieval in an Audio Database," Proc. 3rd ACM Conf. Multimedia (Multimedia '95), ACM Press, 1995, pp. 231–236.

[7] F. F. Kuo and M. K. Shan, "Looking for New, Not Known Music Only: Music Retrieval by Melody Style," Proc. 4th ACM/IEEE-CS Joint Conf. Digital Libraries, (JCDL '04), ACM Press, 2004, pp. 243–251.

[8] P. Ciuha, B. Klemenc and F. solina, "Visualization of concurrent tones in music with colours" Proceedings of the ACM MM '10 2010, pp. 1677-1680.

[9] S. Craig, "Harmonic Visualizations of Tonal Music," Proc. Int. Computer Music Conf., 2001, pp. 423–430.

[10] K. Peacock, "Synesthetic perception: Alexander Scriabin's color hearing," Music Percep. vol. 2, no. 4, 1985, pp. 483–506.

[11] S. Imai, S. Kurabayashi, and Y. Kiyoki, "A Music Database System with Content Analysis and Visualization Mechanisms," Proc. IASTED DIMS 2008, pp. 455–460.

 ISBN: 978-1-61208-280-6

Synote Discussion

Extending Synote to support threaded discussions synchronised with recorded videos

Mike Wald, Yunjia Li, E.A. Draffan, James Brierley, Alyona Ivanova, Robert Streeting, Matthew Tucker
ECS
University of Southampton
UK
m.wald@soton.ac.uk

***Abstract*—Synote Discussion has been developed as an accessible cross device and cross browser HTML5 web-based collaborative replay, annotation and discussion extension of the award winning open source Synote which has since 2008 made web-based recordings easier to access, search, manage, and exploit for learners, teachers and others. While Synote enables users to create comments in 'Synmarks' synchronized with any point in a recording it does not support users to comment on these Synmarks in a discussion thread. Synote Discussion supports commenting on Synmarks stored as discussions in its own database and published as Linked data so they are available for Synote or other systems to use. This paper explains the requirements and design of Synote Discussion, presents the results of a usability study and summarises conclusions and future planned work.**

Keywords- video; synchronise; comment; discussion

I. INTRODUCTION

Synote [1] overcomes the problem that while users can easily bookmark, search, link to, or tag the WHOLE of a recording available on the web they cannot easily find, or associate their notes or resources with, PART of that recording [2]. As an analogy, users would clearly find a textbook difficult to use if it had no contents page, index or page numbers. Synote can use speech recognition to synchronise audio or video recordings of lectures or pre-recorded teaching material with a transcript, slides and images and student or teacher created notes. Synote won the 2009 EUNIS International E-learning Award [3] [4] and 2011 Times Higher Education Outstanding ICT Initiative of the Year award [5]. The system is unique as it is free to use, automatically or manually creates and synchronises transcriptions, allows teachers and students to create real time synchronised notes or tags and facilitates the capture and replay of recordings stored anywhere on the web in a wide range of media formats and browsers. Synote has been developed and evaluated with the involvement of users and with the support of JISC [6] and Net4Voice [7]. Figure 1 shows the Synote player interface. The technical aspects of the system, including the Grails Framework and the Hypermedia Model used, have been explained in detail elsewhere [8]. The synchronised bookmarks, containing notes, tags and links are called Synmarks (Figure 2). When the recording is replayed the currently spoken words are shown highlighted in the transcript. Selecting a Synmark, transcript word or Slide/Image moves the recording to the corresponding synchronised time. The provision of text captions and images synchronized with audio and video enables all their communication qualities and strengths to be available as appropriate for different contexts, content, tasks, learning styles, learning preferences and learning differences. Text can reduce the memory demands of spoken language; speech can better express subtle emotions; while images can communicate moods, relationships and complex information holistically. Synote's synchronised transcripts enable the recordings to be searched while also helping support non-native speakers (e.g., international students) and deaf and hearing impaired students understand the spoken text. The use of text descriptions and annotations of video or images help blind or visually impaired students understand the visually presented information. So that students do not need to retype handwritten notes they had taken in class into Synote after the recording, notes taken live in class on mobile phones or laptops using Twitter [9][10] can

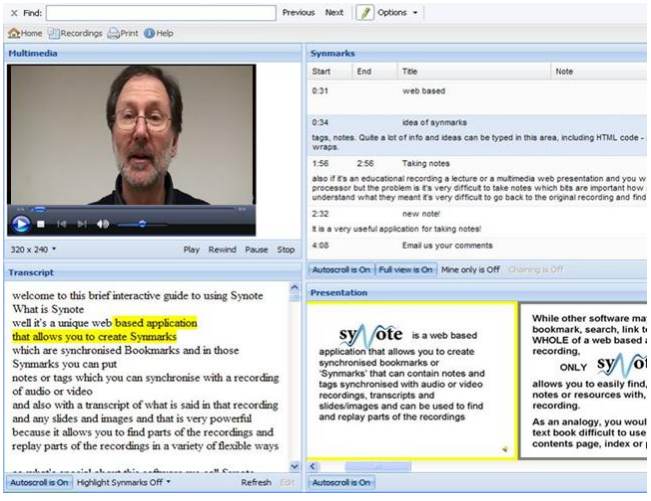


Figure 1. Synote Player Interface

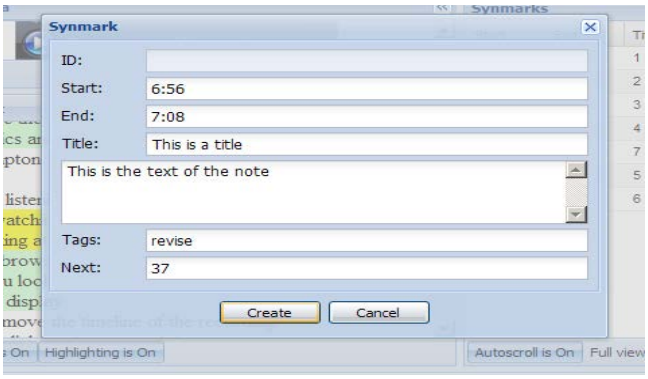


Figure 2. Synote Synmark Creation

be automatically uploaded into Synote. Synote builds on 12 years work on the use of speech recognition for learning in collaboration with IBM, and the international Liberated Learning Consortium [11] [12]. The integration of the speaker independent IBM Hosted Transcription System with Synote has simplified the process of transcription giving word error rates of between 15% - 30% for UK speakers using headset microphones. This compares well with the National Institutes of Standards (NIST) Speech Group reported WER of 28% for individual head mounted microphones in lectures [13].

While most UK students now carry mobile devices capable of replaying Internet video, the majority of these devices cannot replay Synote's accessible, searchable, annotated recordings as Synote was created in 2008 when few students had phones or tablets capable of replaying these videos. Synote Mobile [14] (Figure 3) was therefore developed as an HTML5 responsive design web application capable of working in the majority of browsers on mobile devices running an Android, Windows, Apple iOS system

Figure 3. Comparing Galaxy and iPad presentation of Synote Mobile

that can use YouTube's captioning and transcription service to allow for timed stamped data to be annotated and shared with others [15]. The fall forward approach of MediaElement.js [16] means that should the HTML5 player fail on the device Synote Mobile will present the user with Silverlight or Flash and vice versa in a fall back situation. This solution appears to work with most browsers despite the lack of player access alongside the transcription on smaller mobile devices. It allows for accessibility with captions and transcriptions and provides the user with a way of interacting with others whilst working with video and audio files. This allows for video captured lectures to be not only more accessible to those who have hearing impairments but also allows all students to go back over content in a way that may suit their learning preferences whether they are in the university, at home or when travelling.

Video sharing sites such as YouTube (RX) allow users to discuss the whole video using comments but these discussions cannot be linked to a particular part of the video

Neither Synote nor Synote mobile support threaded discussions as their Synmarks are annotations of the recording timeline. Synote discussion was therefore developed to enable students to have a discussion about a topic raised in the recording in such a way that the discussion is linked to the particular part of the recording being discussed. Section II explains the requirements and design, Section III presents the usability test results while Section IV summarises the conclusions and future work.

II. SYNOTE DISCUSSION REQUIREMENTS & DESIGN

In order to be able to rapidly prototype a system to experiment with the best way of achieving the aims it was decided that rather than redesign Synote's database to allow for this new form of discussion annotation a new database would be created to hold these discussions and the users' threads and comments. In order to ease the integration of Synote Discussion with the original Synote, the comments are further published as linked data using Resource Description Framework (RDF) [17] Key requirements included:

- view Synote presentations as slides or image snapshots of video and an associated transcript with a link to the video on Synote.
- view a list of Synmarks relating to a particular time of the presentation and a list of comments for each Synmark
- add Synmarks and add, edit or delete comments to Synmarks.
- notifications on comments being posted on Synmarks and navigate directly to those Synmarks
- export the discussions as linked data so that it could be accessed and reused by other applications, especially the original Synote
- support the main mobile devices, web browsers and screen sizes in both portrait and landscape modes.

 ISBN: 978-1-61208-280-6

The application was designed to be consistent so none of the features become hidden or removed on different screen sized devices with each page having the same base design and a similar layout for content. Much of the content displayed is interactive with components designed to look 'selectable' and interactive to makes it easier for the user to intuitively navigate the website.

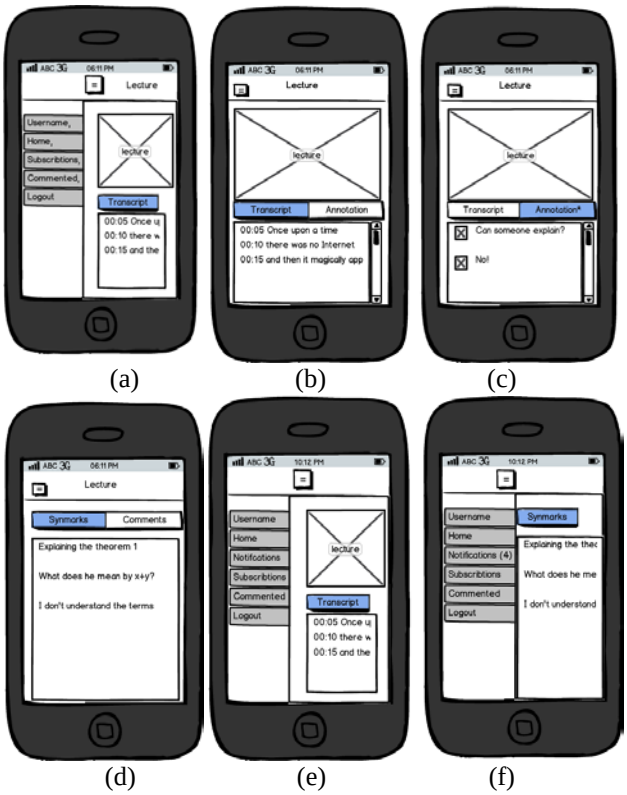


Figure 4. Wireframe sketches of the design.

Figure 4 shows sketches of the designs and this section provides explanations behind the design decisions. Figure 4a shows the menu of the system, which is hidden until the user clicks on the menu button (located in the top left corner of the screen). When the button is selected, the main screen is shifted to the right and the menu bar is shown. This design has been inspired by a number of extremely popular web applications used by the target audience of students, including Facebook Mobile [18], which has a similar sliding menu feature. These menu indicators have become a mobile standard, with this design approach being featured in most major utility mobile applications, not just Facebook.

Figure 4b demonstrates the design of the main detailed view of a presentation in the application. This shows the transcript - a feature taken from Synote - in the bottom half of the screen. As the presentation's slides are being changed, the transcript is kept synchronised with the displayed slide so it is easier for users to follow the lecture.

Figure 4c also shows the main detailed view of a presentation, however, this time in the 'Annotations' tab. This tab holds the Synmarks and Discussions (the top layer of threads). Like the transcript, the Synmarks are synchronised for every time frame (slide) of the lecture. The Synmark's appearance has been made visibly obvious to show that a Synmark is 'clickable' for the user to be redirected to the comments page for that Synmark (not shown in any of the wireframe figures). The page is where a user can view a list of their own Synmarks that they have started in the first tab, and their own comments in the other tab. Every Synmark and Comment in the list is 'selectable,' and directs the user to the presentation page that the content is related to and the Synmark within that presentation.

A very similar design is used for Subscriptions and the users Notifications pages (Fig, 4d & 4e) If the user has any notifications, these are highlighted to the user in the sliding menu, with the number of notifications is in brackets in the menu.

A. *Discussions*

In the system, users can make two levels of annotations to a presentation, similar to the way an online forum works. At the top level are"discussions", which are associated with part of a presentation. These are considered to be equivalent to Synmarks, which are comments on the presentation made in Synote. In order for the discussion to relate to a certain presentation, it stores a presentation ID. However, a discussion should only relate to a certain section of the presentation, so that it can be displayed when the user is viewing that part of the presentation. In order to allow this, a start time and optional end time are stored with the discussion. Since each presentation is split up into a number of sections with their own IDs, this could have been implemented by storing the section ID rather than the presentation ID. However, by implementing it with a start and end time, this allows a discussion to relate to multiple sections, thereby implementing a many-to-many relationship between sections and discussions without having to represent this in the database. Discussions also store an author ID and a timestamp, so that the application can determine when the discussion was created for ordering purposes and who created it.

B. *Comments*

Comments are the second level of annotations, and can be posted in relation to either existing discussions or Synmarks. Since these two types of comments have to store different IDs, the database allows for this by implementing these two types as two different objects, which inherit from a generic Comment object. Therefore, three tables are required: a table for Synmark comments, which stores a generic comment ID and a synmark ID, a table for Discussion comments, which is the same but stores a discussion ID, and finally a generic comment table, which stores the content of the comment, author and various other data about the comment. Rather than allowing users to actually delete comments completely, a deleted field is included instead, which allows a comment to be marked as deleted. This is so that a placeholder can be put in place of the comment so that users can see it is deleted.

 ISBN: 978-1-61208-280-6

C. Subscriptions

Users can subscribe to presentations, Synmarks and discussions to receive notifications when other users comment. Since subscriptions are allowed on all these different entities, a similar inheritance structure to that described for comments is required to make this possible. Even though the general subscription entity only contains a user ID and a primary key, it was decided that the inheritance was still necessary in case it became apparent that more fields were required in the general entity.

D. Notifications

Users need to be notified when either a discussion is posted to a presentation they are subscribed to or a comment is posted to a discussion/Synmark they are subscribed to. Therefore, there are different notifications related to discussions and to comments; this means that inheritance is used once again. The user navigates across the website using a sidebar (Figure 5) that appears on the left of the screen when a button is clicked. The sidebar also displays an indicator showing the number of notifications a user currently has. By only appearing when the user wants to navigate to a new page, this allows the front end to fully utilise the limited space available on a mobile device. The sidebar has its own template which is included in the other templates, allowing it to be changed easily and ensuring it remains consistent across the website.

E. Navigation and Display

The front end uses a mobile-optimised navigation system where, when the user tries to navigate to a new page, the front end fetches the information from the back end and creates it in a page hidden from the user, which then slides into view upon completion, replacing the original page. A loading animation is presented to the user while the page is loaded (Figure 6), and notifies the user if an error has occurred that has prevented the new page from being loaded correctly.

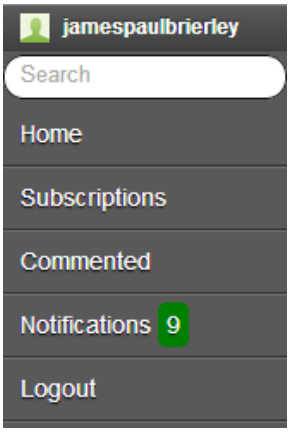


Figure 5. Screenshot of the menu sidebar bar..

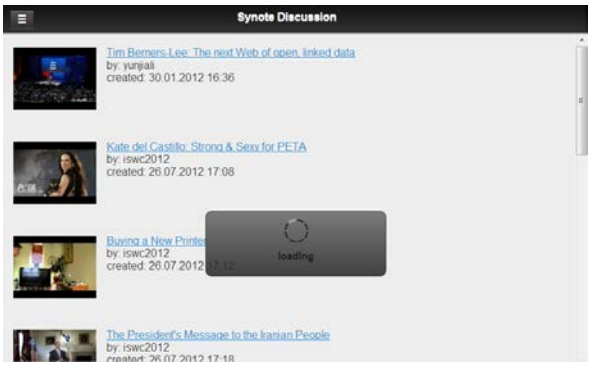


Figure 6. Screenshot showing loading the new page, the menu button can be seen in the top left of the menu sidebar bar.

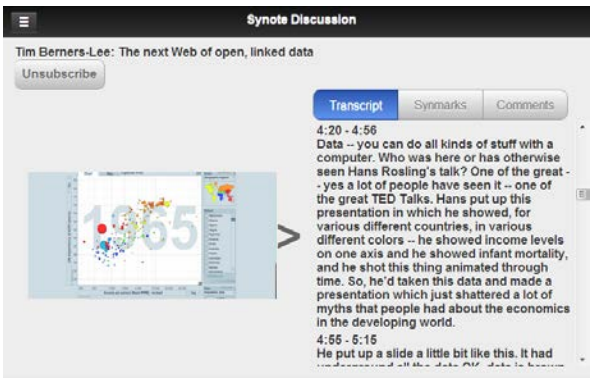


Figure 7. Screenshots showing the presentations displayed in landscape mode.

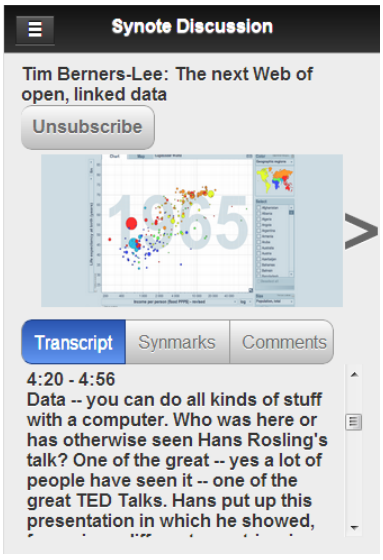


Figure 8. Sketches Screenshots showing the presentations displayed in portrait mode.

The main page of the website, the details page, which displays the presentation has two different display modes, one designed for a landscape aspect ratio (Figure 7) and one

designed for a portrait aspect ratio (Figure 8). This is to allow the website to be displayed in an optimal fashion when a user rotates their mobile device. The presentations are designed to be displayed in as responsive a way as possible. Switching between slides will automatically scroll the transcript to the appropriate section for that slide, and selecting a Synmark will automatically scroll to the image that applies to that Synmark. Users can subscribe to presentations and Synmarks, allowing them to receive notifications when a new Synmark or comment is added. There is also a subscriptions page, where the user can see a list of all their current subscriptions, and remove or navigate to each individual subscription. There is a page where users can view a list of Synmarks and comments that they themselves have added and there is a page where users can view all their notifications, with options to remove individual notifications or clear all notifications.

III. USABILITY TESTS

Usability testing involved 10 university students, the target group for the system, participating remotely. The participants were asked to test the application on any mobile, tablet or desktop web browser of their choosing. They were given directions as to where the application was located and were asked to use it freely for as long as they needed to get acquainted with it. Following this, they were each asked to perform a range of specified actions/tasks. After completing the tasks, they were each asked several questions about their experiences and were asked to give some feedback about the application.

Tasks they were asked to perform were: 1. Login; 2. Explore the home page and navigate to a presentation; 3. Navigate through the presentation and leave a comment on one of the slides; 4. Create their own discussion; 5. Subscribe to that presentation; 6. Subscribe to one of the discussions and leave a comment there; 7. Someone has left a comment on their discussion, how would they quickly find out about it?; 8. From the subscriptions page, quickly navigate to one of the discussions they are subscribed to and unsubscribe from it; 9. Navigate to one of the discussions they have left a comment on; 10. Try to edit and delete some of their comments

IV. USABILITY TEST RESULTS

Most of the users said that initially they were not sure what the purpose of the application was. After they logged in and browsed around, it was slightly clearer that the application was a media store, although it was not clear what types of the media were stored (e.g., whether it was YouTube entertainment videos, or educational recordings from lectures). This may have been due to some of the test content used on Synote being unrelated to educational material. Some of the participants also expressed their hesitation with using their external authentication details, even though they had been explained how to revoke the access from the application to their details later on. All the participants said that it would have been more convenient for them to at least to have an option of being able to register with the system separately and hence not to give away their external details. Some participants found it was irritating that on some of their devices they encountered loading errors and the only fix offered was reloading the application. Although it did not cause any major test disturbance, it was concluded that this issue could put off the potential users from using the application. Most of the participants did not understand (until they were told) what the Synmarks were and why they could not add comments to a presentation without selecting a Synmark related to a slide. It was concluded that because most of the participants were not familiar with the original Synote system and terminology, they could not have been expected to understand this initially. The participants also expressed a slight confusion with the fact that the Synmarks were moving the presentation slides, although all of the participants agreed that synchronizing transcript and Synmarks according to the time frame of the related slide was one of the most useful features. However, according to the users the confusing comments tab when no Synmark is selected, should not appear. All of the participants agreed that because of the lack of distracting information, it was easier to perform the tasks and the application was very straightforward and intuitive to use. All of them also concluded that because the application has a very typical mobile platform look, it was very easy to learn how to navigate around whilst performing the tasks. It was agreed that none of the features were hidden from the view or were awkward to uncover in order to complete the test. One of the participants did, however, find it slightly irritating to keep checking the (hidden) menu bar for the notifications. It was suggested that the notifications should constantly be visible to a user, no matter whether the sliding menu was in its hidden or visible state. All of the participants gave positive feedback on the design layout and overall design consistency. It was agreed that the layout of the items and colours was generally reasonable and intuitive. It was admitted, however, that because the application was designed to be mainly served as a mobile application, its layout and components are more pleasant to the eye on tablets and mobiles rather than desktop browsers. The application also received positive feedback on how visually”adaptable” it was as the screen size and platform the participant used did not have an effect. Most of the participants also agreed on the usefulness of such an application and its features in their studies, although some of them said that the unfamiliarity with the original Synote would put them off. Almost all of the participants agreed that there would not be any situations when the application would be hard or impossible to use. Only one participant suggested a weak point was the requirement to have a constant Internet connection in order to access and manipulate the data. There could be potential issues with the connection availability due to the content being available only online. All of the test participants concluded that the reason they had some difficulties with understanding what the system was for and what they were to do to complete the set tasks, was that there were no help messages provided for their learning. Participants also suggested there should be different user

 ISBN: 978-1-61208-280-6

groups for the resource owners and so separate the lecturers/uploaders from the student resource users. In particular that feature would be useful during the discussions as it would be easier to distinguish lecturers from the students.

Overall the users had a positive experience with the system, referring to it as friendly, straightforward and, best of all, compatible with several platforms. The feedback proved useful and influenced the development of the system.

V. CONCLUSION AND FUTURE WORK

Synote Discussion has been developed as an accessible cross device and cross browser HTML5 web-based collaborative replay, annotation and discussion extension of the award winning open source Synote which has since 2008 made web-based recordings easier to access, search, manage, and exploit for learners, teachers and others. While Synote enables users to create comments in 'Synmarks' synchronized with any point in a recording it does not support users to comment on these Synmarks in a discussion thread as Synmarks are annotations of the recording timeline. Synote discussion was therefore developed to enable students to have a discussion about a topic raised in the recording in such a way that the discussion is linked to the particular part of the recording being discussed.

Synote Discussion supports commenting on Synmarks stored as discussions in its own database and published as Linked data so they are available for Synote or other systems to use. Synote Discussion provides an extension to Synote with discussion commenting and Facebook style notifications that work effectively on iOS and Android mobile operating systems and common mobile web browsers. The key requirements were met to enable Synote's data to be displayed and augmented with discussions and comments while Synote Discussion's stored discussion threads and comments can be accessed as open linked data.

The usability tests showed that once users understood the purpose of the application, they found all of the features easily to use but accessing Synote Discussion through a link from Synote would make the relationship between the two applications clearer.

It is planned to introduce this link from Synote and undertake further trials of Synote discussion in real classes with a larger number of users and including a wider range of mobile, tablet or desktop web browsers.

REFERENCES

[1] M. Wald, et. al, "Synchronised Annotation of Multimedia," IEEE International Conference on Advanced Learning Technologies, 2009, 594-596.

[2] S. Whittaker, P. Hyland, and M. Wiley, "Filochat handwritten notes provide access to recorded conversations," Proceedings of CHI, 1994, 271-277.

[3] http://www.ecs.soton.ac.uk/about/news/2598 last accessed 24/03/2013

[4] http://www.eunis.org/index.php/en/jens-doerup-award.html last accessed 24/03/2013

[5] http://www.ecs.soton.ac.uk/about/news/3874 last accessed 24/03/2013

[6] http://www.jisc.ac.uk last accessed 24/03/2013

[7] http://spazivirtuali.unibo.it/net4voice/default.aspx last accessed 24/03/2013

[8] Y. Li, et. al, "Synote: development of a Web-based tool for synchronized annotations," New Review of Hypermedia and Multimedia, 2011, pp. 1-18.

[9] http://twitter.com last accessed 24/03/2013 last accessed 24/03/2013

[10] http://www.ecs.soton.ac.uk/about/news/2812 last accessed 24/03/2013

[11] D. Leitch, and T. MacMillan, "Liberated Learning Initiative Innovative Technology and Inclusion: Current Issues and Future Directions for Liberated Learning Research," Saint Mary's University, Nova Scotia, 2003 http://liberatedlearning.com/wp-content/uploads/2011/05/2003_Year_IV_Research_Report.pdf last accessed 24/03/2013

[12] M. Wald, and K. Bain, "Enhancing the Usability of Real-Time Speech Recognition Captioning through Personalised Displays and Real-Time Multiple Speaker Editing and Annotation," In Proceedings of HCI International 2007: 12th International Conference on Human-Computer Interaction, Beijing, 2007, pp 446-452.

[13] J. Fiscus, N. Radde, J. Garofolo, A. Le, J. Ajot, and C. Laprun, " The Rich Transcription," Spring Meeting Recognition Evaluation, National Institute Of Standards and Technology, 2005

[14] http://linkeddata.synote.org/ last accessed 24/03/2013

[15] http://linkeddata.synote.org/synote/recording/replay/52593 last accessed 24/03/2013

[16] http://mediaelementjs.com last accessed 24/03/2013

[17] T. Berners-Lee, " Linked Data - Design Issues." 2006 http://www.w3.org/DesignIssues/LinkedData.html last accessed 24/03/2013

[18] http://m.facebook.com last accessed 24/03/2013

 ISBN: 978-1-61208-280-6

A Model-Driven Approach for Service Oriented Web 2.0 Mashup Development

José Luis Herrero, Pablo Carmona, Fabiola Lucio
Department of Computer and Telematics Systems Engineering
University of Extremadura
Avda. de Elvas s/n, 06006, Badajoz, Spain
{jherrero,pablo,flucio}@unex.es

***Abstract*— Mashup applications are composed by data or functionality extracted from different sources. With the evolution of web 2.0 and the appearance of AJAX technology and the service-oriented architecture, a new breed of mashup applications for the web has emerged. However, software engineers have to deal with the heterogeneous composition of mashup sources, which increases software development cost and complexity. Therefore, it becomes essential to boost a software development approach that can attenuate these problems. This is the reason why we propose in this paper a model-driven and service-oriented architecture for developing mashup applications. Towards this end, the following tasks have been developed: first a new mashup profile extends UML and introduces mashup concepts at design level, and second, a set of transformation rules has been defined with the aim of generating code semi-automatically. These rules have been classified according to the type of the element (web application, mashup or web service). Finally, a transformation tool parses a UML model, identifies mashup elements, and according to the specified set of rules, generates code.**

Keyword: Mashup, Model-driven architecture, web services, web applications.

I. INTRODUCTION

With the appearance Web 2.0 paradigm and the introduction of new technologies such as Asynchronous JavaScript and XML (AJAX) [1] and web services, new types of applications for the web have emerged. Under the umbrella of this new trend, Rich Internet Applications (RIA) [2] has evolved, and a high degree of interactivity and complexity is achieved by this type of applications. One of the most interesting type of applications that have gained much attention in the Web 2.0 community, is mashups. During the last few years, different types of mashups have been defined: on the one hand, data mashups have the ability to produce new information combining data from different sources, and on the other hand, functional mashups are composed by mashup components that can be assembled and combined in order to build the final application.

The service-oriented architecture (SOA) is defined as ‘‘a set of components, which can be invoked, and whose interface descriptions can be published and discovered’’ [3]. One implementation of SOA applications is made possible through the realization of Web Services, which are implemented in eXtended Markup Language (XML), and described through the Web Services Description Language (WSDL), while the simple object access protocol (SOAP) is the main communication protocol adopted.

A mashup is a program that manipulates and composes existing data sources or functionality to create a new piece of data or service that can be plugged into a web application [4]. SOA provides a solid foundation for mashup development. However we argue that the underneath technology that supports mashup applications requires software engineers to locate and combine mashup sources, which implies an increase in the complexity degree, and in particular in the development costs of this type of applications.

In order to solve these problems, we propose in this paper a Model Driven Architecture (MDA) [5]. It simplifies modeling, design, implementation, and integration of applications by defining software mainly at the model level. The primary goals of MDA are portability, interoperability, and reusability through architectural separation of concerns [6], making product development more cost efficient by increasing automation in software development [7].

The objective of this paper is to propose a model-driven architecture to develop mashup applications, and this task has been achieved at the following levels: first, a new profile that includes specialized concepts from the mashup domain has been defined, and then, a transformation model generates mashup applications from the UML profile. This paper is organized as follows: first, Section II explains the motivation and the background of this work. Then, a UML mashup profile is proposed in Section III, and also an example is presented. Next, Section IV describes the transformation model to generate mashup applications from a UML design, and finally, conclusions are clarified in Section V.

II. MOTIVATION AND RELATED WORKS

Two different taxonomies can be mentioned when classifying mashups applications; the first one is focused on the place the composition mechanism takes place [8], and the other one studies the type of the combined elements [9].

This classification brings us the motivation to propose a novel architecture that captures the essence of both alternatives. On the one hand, mashups can be composed at server or client side, while on the other hand, the composition can be focused on data or functional components. As far as we know, this is a novel approach to deal with mashup applications development.

 ISBN: 978-1-61208-280-6

A. Mashup background

Different tools have been proposed to build mashup applications for the web. Dapper [10] is a drag and drop tool that allows users to select contents in several web pages that will be composed in order to generate a new representation. Yahoo Pipes [11] is a composition tool to aggregate and manipulate mashup content from the web. This tool provides a visual editor to create pipes from different sources and provides rules to compose the content. DERI [12] is inspired by Yahoo's Pipes, and proposes an engine and graphical environment for general web data transformations and mashup. Serena Business Manager [13] contains a visual workflow editor to define mashups and presents an online marketplace where mashups can be exchanged. And finally, IBM Mashups Center [14] is a mashup platform that supports rapid assembly of dynamic web applications, enabling the creation, sharing, and discovery of reusable application building blocks that can be easily assembled into new applications.

Trends in the mashup community are currently working in different areas, Meditskos et al. [15] proposes an approach for developing mashups with semantic mashup discovery capabilities has been proposed, and an extension of OWL-S advertisements has been defined. A novel service oriented architecture is presented in [16] which addresses reusability and integration needs for building mashup applications, identifying the essential architecture patterns for designing mashups. A different work [17] approaches to represent domain concepts at the mashup composition, and defines an architecture to assist experts in the process of introducing domain concepts in the composition mashup level. Another interesting area [18] studies the privacy problem, which deals with the dynamic data integration from different mashup sources in the presence of privacy concerns, and proposes a service-oriented architecture for privacy-preserving data mashup.

B. MDA background

In the field of MDA, there is a consensus about the benefits that this technology offers for software development: a reduction of sensitivity to the inevitable changes that affect a software system [19], a reduction of cost and complexity [20], and an increase of abstraction [21]. An interesting analysis about the existing problems in the field of web engineering and how they can be solved by model-driven development approaches is presented in [22], which identifies the problems encountered in the development process of web applications such as their dependence on the HTTP protocol, compatibility issues due to the heterogeneity of web browsers, and the lack of performance because of the increase in the latency degree

Different proposals extend web engineering methods for developing web applications. Fraternali et al. [23] present a survey of existing web engineering methods to develop this type of applications. The Object-Oriented Hypermedia Design Model (OOHDM) [24] uses abstraction and composition mechanisms in an object-oriented framework to, on one hand, allow a concise description of complex information items, and on the other hand, allow the specification of complex navigation patterns and interface transformations. The RUX-Model [25] is a representational model that offers a method for engineering the adaptation of legacy model-based Web 1.0 applications to Web 2.0 UI expectations. An extension of this model is proposed in [26] where a model-driven approach to web application development by combining the UML based Web Engineering (UWE) with the RUX-Method is defined.

The combination of these two trends has been studied in [27]. The article proposes a model-driven mashup development (MDMD) as an approach to develop mashups according to flexible framework (PLEF-Ext) for end-user. A Service-Oriented Model Driven Architecture (ODSOMDA) approach has also been proposed in [28], which involves adding service oriented architecture (SOA) elements into a model-driven architecture to facilitate the construction of mashup applications.

III. BUILDING MASHUPS WITH A MDA APPROACH

MDA comprises of three main layers: a) the Computation Independent Model (CIM) is the top layer and represents an abstract model of the system, abstracting from technical details, b) the Platform Independent Model (PIM) defines the conceptual model based on visual diagrams, use-case diagrams and metadata, and c) the Platform Specific Model (PSM) that represents the system from a specific implementation platform viewpoint. In order to achieve an implementation of the system, the PIM must be transformed into PSM, and with this aim, an automatic code generator must guide this process. The most important advantage of this approach is that PSMs can be automatic generated from a PIM model according to a set of transformation rules.

This paper proposes a MDA approach to develop mashup applications for the web. With this aim, a new UML profile is proposed (Figure 2), and a set of a transformation rules has been developed in order to generate mashup applications according to the WCF framework [29].

A. Mashup profile

The main element of this profile is the <<Mashup Application>> stereotype that represents a mashup application, which includes the name and composition type (client or server). A mashup application can be attached to a web application, represented by the <<Web Application>> stereotype. A mashup resource (<<Mashup resource>>) specifies a data or software element that can be combined in order to create new information or add new functionality. Mashup resources are classified as an enterprise mashup (<<Enterprise mashup>>) or data mashup (<<Data Mashup>>). <<Mashup component>>, <<web API>> and <<Widget>> stereotypes represent the different types of enterprise mashups.

Data mashups can be categorized according to data integration patterns: the <<Mashup Pipe>> stereotype defines a set of pipes or filters that must be applied to the information in order to obtain the final output.

 ISBN: 978-1-61208-280-6

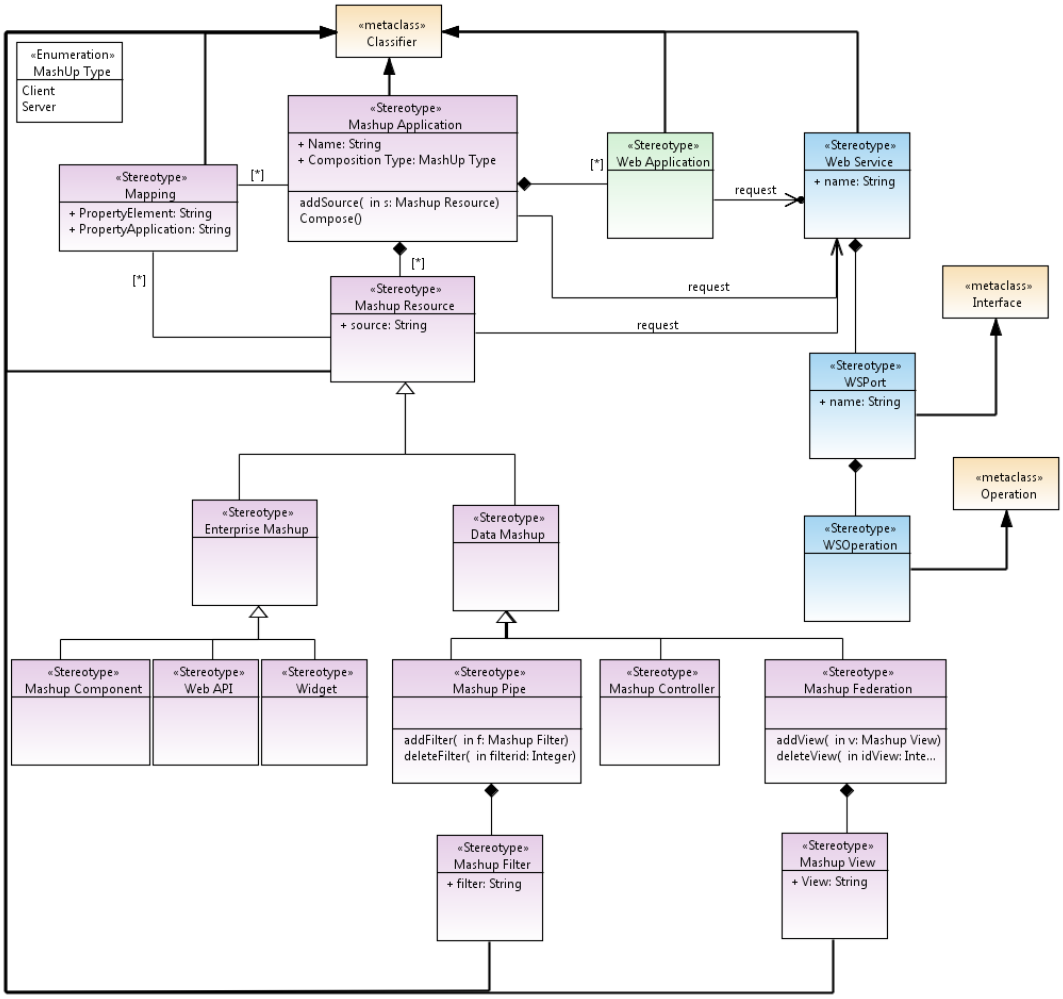


Figure 1. UML mashup profile

The <<Mashup DataFederation>> stereotype specifies different views of the same data, while the <<Mashup Controller>> stereotype describes how the data is rendered. The mapping between a mashup application and its sources is specified by the <<Mapping>> stereotype.

Finally, services are referenced by the <<Web Service>>, <<WSPort>> and <<WSOperation>> stereotypes according to a previous proposal [30], but avoiding unnecessary stereotypes in order to provide a more compact profile.

B. *An illustrative scenario*

In order to test the proposed profile, the following scenario is suggested: an ecommerce web portal (Figure 3) allows clients to buy products that can be offered by different providers. Additionally, a set of tools are supported in order to provide useful utilities (currency conversions, a calendar tool and a shopping cart). A payment system is defined and connections with bank payment services are also supplied. Clients, providers and banks are connected with the ecommerce portal using web services technology.

IV. MODEL TRANSFORMATIONS

One of the keys of MDA is the capacity of defining transformations from higher-level models to platform specific models guided by a set of transformation rules. With this aim, a generation tool is proposed in this section. This tool is based on Eclipse platform (Eclipse Juno version), and different plugins have been used in order to support UML development (Papyrus Project) and code generation (Acceleo Project). A library has also been developed with the aim of requesting elements from a UML design, and extract information about the model.

The transformation process parses a UML model, next identifies all the elements tagged with the new stereotypes defined in the mashup profile, then extracts all the information about them, and finally generates the specific code according to the Web Component Framework (WCF).

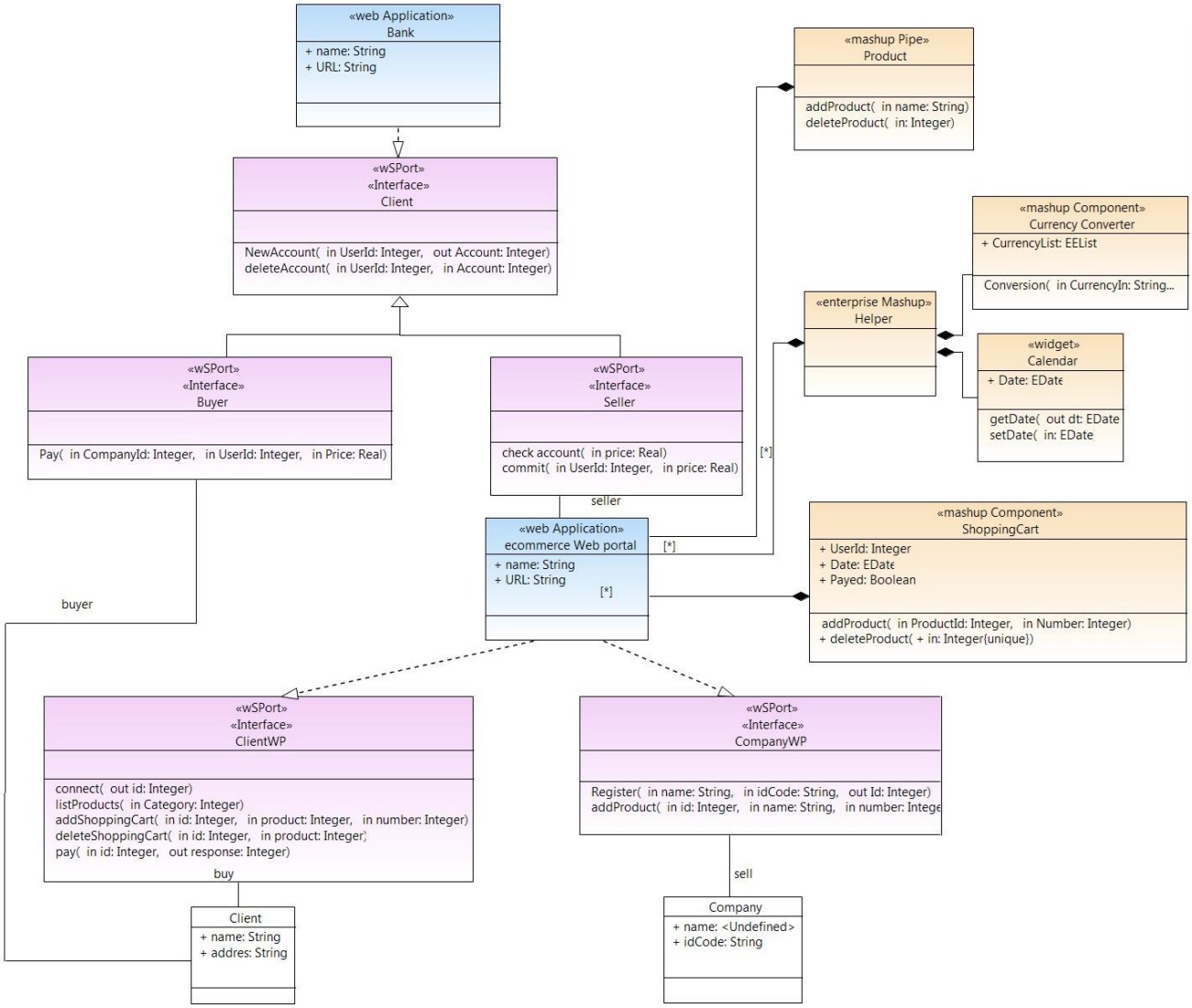


Figure 2. Ecommerce web portal.

In order to perform the transformation, the following definitions have been assumed.

TABLE I. DEFINITIONS

Definition 1	Let *M* be the model designed according to the mashup profile
Definition 2	Let *Cm* be an element extracted from the model.
Definition 3	Let *As* be an Association between two *Cm* elements,.
Definition 4	Let *End* be a target element of an Association *As*.
Definition 5	Let *Ma* be a model element tagged as a <<Mashup Application>>.
Definition 6	Let *WSP* be a model element tagged as a <<WSPort>>.

The set of rules proposed are classified according to the type of element generated: web application, mashup and web service.

1) Web Application rule

This rule searches each *Cm* in the model *M*, tagged as a "Web Application", looks for its associations and checks if a Mashup application is attached. In this case a mashup application is created.

Input: a UML model (M)
Output: generates the definition of a Web application as a collection of mashup applications

```
1  for each Cm ϵ Elements(M) do
2    if (getStereotype(Cm)="Web Application")
3      for each As ϵ Associations(Cm) do
4          if (typeof(As)="Aggregation")
5          for each End ϵ AssociationEnds(As) do
6            if (getStereotype(End)="Mashup Application")
```

```
7                out (“createMashupApplication”)
8                out (“ (“+End.getAttribute(“name”)+”,”)
9                out (End.getAttribute(“type”)+”)”)
10               generateMashupResources(End)
11             end if
12           end for
13       end for
14     end if
15  end for
```

2) Mashup rule

First, the rule checks every mashup resource attached to a mashup application *Ma.* Next, according to the type a specific mashup resource is generated (data or enterprise).

Name: generateMashupResources
Input: mashup Application (Ma)
Output: generates the definition of mashups resources.

```
1     for each As ϵ Associations(Ma) do
2       if (typeof(As)=”Aggregation”)
3         for each End ϵ AssociationEnds(As) do
4           if (End .isSubtypeof(”Mashup Resource”)
5             if (End .isSubtypeof(”Enterprise Mashup”)
6               out(CreateEnterpriseMashup+”(”)
7             else
8               if (End .isSubtypeof(”Data Mashup”)
9                   out(CreateDataMashup+”(”)
10            end if
11          end if
12          out(End.getAttribute(“name”)+”,”)
13          out(End.getAttribute(“Composition Type”)+”,”)
14          out(End.getAttribute(“source”)+”)”)
15          out (“addsource(‘“+Ma.getAttribute(“name”)+”’,’”+
            End.getAttribute(“name”)+”’)”)
16        end if
17      end for
18    end if
19  end for
```

3) Web Services

The last elements considered in this transformation are web services, which are represented using different stereotypes (Web services, WSPort and WSOperation). The transformation from these elements to the Web Services Description Languag is guided by the following rules:

a) Header and definitions rule

The header and the definition part of the web service are generated by this rule. With this aim, each element tagged with the *Web Service* stereotype is located and every *WSPort* element attached is extracted.

Input: UML model (M)
Output: generates the web service header and definition

```
1   for each Cm ϵ Elements(M) do
2     if (getStereotype(Cm)=”Web Service”)
3       out(“<?xml version=’1.0’ encoding=’UTF-8’?>”)
4       out(“ <definitions name=’”+Cm.getAttribute(“name”)
          +”Service’>”)
5       for each As ϵ Associations(Cm) do
6         if (typeof(As)=”Aggregation”)
7           for each End ϵ AssociationEnds(As) do
8             if (getStereotype(End)=”WSPort”)
9               generateMessages(End)
10              out(“</definitions>”)
11            end if
12          end for
13        end if
14      end for
15    end if
16  end for
```

b) Messages rule

The set of request and response messages are extracted from each *WSOperation* attached to a *WSPort* and the WDSL message part is generated.

Name: generateMessages
Input: WSP is a “WsPort” element
Output: generates web service messages

```
1   for each As ϵ Associations(WSP) do
3     if (typeof(As)=”Aggregation”)
4       for each End ϵ AssociationEnds(As) do
5         if (getStereotype(End)= ”WSOperation”)
6           out(“<message name=”+End.getAttribute(“name”)+
            ”Request>”
7           for each Op ϵ Operation do
8             out(Op.getAttribute(“name”)+”Request”)
9             for each P ϵ Params(Op) do
10              if (P.direction=”in”)
11                out(“<part name=”+P.getAttibute(“name”)+
                  ”type=xs:”+P.getAttribute.type+”’/>”
12              end if
13            end for
14            out(“</message>”)
12            out(Op.getAttribute(“name”)+”Response”)
13            for each P ϵ Params(Op) do
14              if (P.direction=”out”)
15                out(“<part name=”+P.getAttibute(“name”)+
                  ”type=xs:”+P.getAttribute.type+”’/>”
16              end if
17            end for
18        end for
19      end if
20    end for
21  end if
22 end for
```

c) Binding rule

This rule generates the binding part of the web service and describes how the service is bound to a SOAP message protocol.

Name: generateBinding
Input: Ws is a “WsPort” element
Output: generates web service binding

```
1   out(“<binding name=’+Ws.getAttribute(“name”)+”Bind’
    type =’tns:”+Ws.getAttribute(“name”)+”Port’>)
2   out “(<soap:binding style='document' transport='http://schemas.
    xmlsoap.org/soap/http'/>”)
2   for each As ϵ Associations(End) do
3     if (typeof(As)=”Aggregation”)
4       for each End ϵ AssociationEnds(As) do
5         if (getStereotype(End)= ”WSOperation”)
6           for each Op ϵ Operation do
8             out <operation name=’”+Op.name/]+”’>”
9             out(“<input><soap:body use=’literal’/></input>”)
10            out(“<output><soap:body use=’literal’/></input>”)
12          end for
```

```
13          end if
14        end for
15      end if
16  end for
17  out("</binding")
18  out("<service name="+Ws.getAttribute("name") +"Service'>")
19  out("<port binding='tns:"+ Ws.getAttribute("name") +"Bind")
20  out("name="+Ws.getAttribute("name") +"Port")
21  out(</port></service>")
```

V. CONCLUSIONS AND FUTURE WORKS

Mashup applications provide the capacity of creating new elements by composing existing resources, and this is the reason why they are being widely adopted in the web 2.0 community.

The presented work tries to enrich mashups with the definition of a model-driven approach that provides new advantages in the development process of this type of applications. The main purpose of this proposal is making mashup development more cost efficient by increasing automation in software development. With this aim, the following task has been performed: a new profile extends UML in order to specify new concepts involved in a mashup application, and a set of transformation rules have been propose with the aim of guiding the transformation process.

Future works will try to study how to increase the performance degree of our approach. With this aim, we will try to incorporate prefetching techniques to download web contents in advance.

REFERENCES

[1] J. Garrett. "Ajax: A new approach to web applications", http://www.adaptivepath.com/publications/essays/archives/000385.php. [retrieved:December, 2012.

[2] L.D.Paulson, "Building rich web applications with Ajax", Computer, vol.38, no.10, 2005, pp. 14-17], doi: 10.1109/MC.2005.330.

[3] W3C, World Wide Web Consortium. http://www.w3c.org. [retrieved:November, 2012]

[4] K. Stolee and S. Elbaum, "Refactoring pipe-like mashups for end-user programmers", 33rd International Conference on Software Engineering (ICSE '11), 2011, pp. 81-90, doi:10.1145/1985793.1985805.

[5] OMG. Model Driven Architecture, http://www.omg.org/mda [retrieved:November, 2012]

[6] J Estefan, "Survey of Model-Based Systems Engineering (MBSE) methodologies",http://www.incose.org/productspubs/pdf/techdata/mttc/mbsemethodogy_survey_2008-0610_revb-jae2.pdf, [retrieved: January, 2013]

[7] S. Teppola, P. Parviainen, and J.Takal, "Challenges in Deployment of Model Driven Development", Fourth International Conference on Software Engineering Advances (ICSEA '09), 2009, pp.15-20, doi: 10.1109/ICSEA.2009.11.

[8] S. Aghaee and C Pautasso, "Mashup development with HTML5", 3rd and 4th International Workshop on Web APIs and Services Mashups. article No. 10, 2010, doi:10.1145/1944999.1945009

[9] P Vrieze, L. Xu, A. Bouguettaya, J. Yang, and J. Chen, "Building enterprise mashups", Future Generation Computer Systems, vol. 27, issue 5,2011, pp.637-642.

[10] Dapper: The Data Mapper. http://www.dapper.net/. [retrieved:January, 2013]

[11] Yahoo Pipes. http://pipes.yahoo.com/pipes/. [retrieved:January, 2013]

[12] DERI Pipes. http://pipes.deri.org/. [retrieved:January, 2013]

[13] Serena Business Manager. http://www.serena.com/index.php/en/products/sbm. [retrieved:January, 2013]

[14] IBM Mashup Center. http://www.ibm.com/developerworks/lotus/products/mashups/. [retrieved:January, 2013]

[15] G. Meditskos and N. Bassiliades, "A combinatory framework of Web 2.0 mashup tools, OWL-S and UDDI", Journal Expert Systems with Applications, volume 38 Issue 6, 2011, pp. 6657-6668, doi:10.1016/j.eswa.2010.11.072

[16] Y. Liu, X. Liang, L. Xu, M. Staples, and L.Zhu, "Composing enterprise mashup components and services using architecture integration patterns". Journal of Systems and Software archive, volume 84, issue 9, 2011, pp. 1436-1446, doi:10.1016/j.jss.2011.01.030

[17] S. Soi and M. Baez, "Domain-specific Mashups: From All to All You Need", Proceedings of the 10th international conference on Current trends in web engineering (ICWE'10), 2010, pp. 384-395.

[18] B.C.M. Fung, T. Trojer, P.C.K. Hung, L. Xiong, and K.Al-Hussaeni, "Service-Oriented Architecture for High-Dimensional Private Data Mashup", IEEE Transactions on Services Computing, 2012, vol. 5, no. 3, pp. 373-386, doi: 10.1109/TSC.2011.13.

[19] C. Atkinson and T. Kühne, "The role of metamodeling in MDA". International Workshop in Software Model Engineering, 2002.

[20] J. Mukerji and J. Miller. MDA Guide version 1.0.1, http://www.omg.org/cgi-bin/doc?omg/2003-06-01, [retrieved: December, 2012]

[21] G. Booch, A.W. Brown, S. Iyengar, J. Rumbaugh, and B Selic, "An MDA manifesto", MDA Journal., 2004. http://www.bptrends.com/publicationfiles/05-04 COL IBM Manifest 20 -Frankel- 3.pdf. [retrieved:January, 2013]

[22] R. Gitzel, A. Korthaus, and M. Schader, "Using established Web Engineering knowledge in model-driven approaches", Science of Computer Programming, 2007, vol. 66, issue 2, pp. 105-124, doi:10.1016/j.scico.2006.09.001.

[23] P. Fraternali, S. Comai, A. Bozzon, and G.T. Carughi, "Engineering rich internet applications with a model-driven approach", Journal ACM Transactions on the Web (TWEB), 2010, vol. 4, issue 2, doi:10.1145/1734200.1734204

[24] D. Schwabe, G. Rossi, and S.D.J. Barbosa, "Developing Hypermedia Applications using OOHDM", Seventh ACM conference on Hypertext, 1996, pp. 116-128.

[25] M. Linaje, J.C. Preciado, and F. Sanchez, "Engineering Rich Internet Application User Interfaces over Legacy Web Models", Journal IEEE Internet Computing archive, volume 11, issue 6, 2007, pp. 53-59.

[26] J.C. Preciado, M. Linaje, R. Morales, F.Sanchez, G. Zhang, C. Kroiβ, and N. Koch, "Designing Rich Internet Applications Combining UWE and RUX-Method", Eighth International Conference on Web Engineering, 2008, pp. 148-154, doi:10.1109/ICWE.2008.26.

[27] M.A. Chatti, M. Jarke, M. Specht, U. Schroeder, and D. Dahl, "Model-Driven Mashup Personal Learning Environments", International Journal of Technology Enhanced Learning, volume 3, issue 1, 2011, pp. 21-39, doi:10.1504/IJTEL.2011.039062.

[28] K .He, Wang, J. Wang, J.Liu, C. Wang, and H. Lu, "On-Demand Service-Oriented MDA Approach for SaaS and Enterprise Mashup Application Development", International Conference on Cloud and Service Computing (CSC), 2012, pp. 96–103.

[29] J.L Herrero, P. Carmona, and F. Lucio, "Web services and web components". Seventh International Conference on Next Generation Web Services Practices, 2011, pp. 164-170.

[30] D.Skogan, R. Groenmo, and I. Solheim, "Web Service Composition in UML", Eighth IEEE International of the Enterprise Distributed Object Computing Conference, 2004, pp. 47-57.

 ISBN: 978-1-61208-280-6

Digital Badges in Informal Learning Environments

Bradley S. Barker
Institute of Agriculture and Natural Resources,
University of Nebraska-Lincoln
Lincoln, NE
e-mail: bbarker@unl.edu

Abstract— **The awarding of digital badges has become pervasive across social media systems. Digital badges are visual representations of individual accomplishments and/or competencies and skills. The Mozilla Foundation is leading an effort to standardize the issuing and collection of digital badges with the release of the Open Badges Infrastructure (OBI). The OBI provides a framework for educational institutions and employers to issue digital badges and for badge earners to share their badges across social media systems. This paper discusses the idea of incorporating digital badges into the National 4-H Recognition Model (USDA, n.d.). 4-H in the United States is a youth development organization that serves nearly 6.5 million youth and is administered by the United States Department of Agriculture (USDA). If done properly, the awarding of digital badges may increase the goal setting habits and motivate young people to join positive youth development organizations like 4-H.**

Keywords-Digital Badges; informal learning environments; 4-H; Robotics

I. INTRODUCTION

A diploma is the ultimate outward sign of learning accomplishment. However, some would argue that such diplomas do not communicate detailed skills and abilities nor do they provide information on those skills learned outside of the classroom for learners [1]. Mozilla, a non-profit software development company, has developed the Open Badges Infrastructure (OBI) for learners to showcase their skills and abilities through digital badges. Researchers and educators are interested in the unique opportunity to apply community-based assessment practices whereby the badge issuer authenticates the skills and abilities of the learners that lead to the awarding of a digital badge. Community-based assessments, unlike formal educational courses, may be done by course developers, peers, the content management system, or even by learners themselves [5]. There is also interest in the role digital badges play in goal setting and motivation to pursue additional academic experiences in out-of-school time learning.

This paper specifically discusses the idea of incorporating the Mozilla Open Badges Infrastructure (OBI) into the National 4-H Recognition Model [1]. The 4-H organization in the United States is a youth development organization that serves nearly 6.5 million youth and is administered by the United States Department of Agriculture (USDA). *Head, hearts, hands,* and *health* are the four Hs in 4-H and represent the values of the organization. If done properly, the awarding of digital badges may increase the goal setting habits and motivate young people to join positive youth development organizations like 4-H. The 4-H program in the United States was originated by the Smith-Level act of 1914, which created the Cooperative Extension System, a unique partnership with the 109 land-grant universities in the United States, the USDA, and the National Institute of Food and Agriculture [2]. Currently 4-H is the largest youth serving organization with about 6.5 million participants and is found in every state and territory of the US. Educational programs are delivered in out-of-school programming, in-school enrichment programs, clubs and camps. 4-H offers a wide variety of science, engineering, technology and applied mathematics educational opportunities – from agricultural and animal sciences to rocketry, robotics, environmental protection and computer science [2]. The rest of this paper is organized as follows. Section II presents the current need for a digital badging system. The specific purpose of digital badges is presented in Section III. Section IV describes the OBI and it's implementation in the 4-H youth development system. Section V contains the conclusions and future work.

II. NEED FOR DIGITAL BADGES

Beyond the formal classroom, there are many opportunities to develop and refine skills and abilities as a young person or as an adult. These experiences include visiting informal learning institutions, participating in informal focused learning programs, and utilizing media to pursue interests [3]. Consider that learning is a lifelong endeavor and often is accomplished through self-interest projects, self-directed experiences, information gathering, and community participation or on-the job experience [4]. While these learning pursuits can be personally fulfilling, often times, employers and educational institutions do not recognize or place value on the skills and abilities developed through informal learning. In addition, with the proliferation of computing technologies along with the interconnectedness of the networked learning environments education has been transformed from the classroom to many web-connected spaces. Digital badges can act as a bridge between formal education and the larger connected learning environments by recognizing skills and competencies across these contexts also known as learning ecosystems.

 ISBN: 978-1-61208-280-6

While Mozilla and others advocate the use of digital badges to document individual's discrete skills and abilities, there are some concerns with awarding badges. One potential consequence of awarding badges is to overemphasize the badge as an external reward for learning thereby reducing the learners' intrinsic motivation to learn [5]. This may lead to the gameification of the system where the end goal is badge collection and not the pursuit of knowledge. Another concern lies in the openness of the badge system. The open badges ecosystem will permit the development of many different badges each with a different degree of rigor behind the issuance of individual badges. With such diversity, educational institutions and employers may not accept the premise that a particular badge truly represent the skills and abilities professed. Finally, a concern for the 4-H system is the reliance on community-based assessment and the responsibility of adult facilitators to evenly apply the badge assessments throughout the system. The possibility that a few educators would make exemptions in the badge assessment requirements could jeopardize the outward legitimacy of the entire system to employers and educational institutions.

III. PURPOSE OF DIGITAL BADGES

Digital badges have a wide range of purposes from social belonging and status to representing individual learning achievement. An example of a social belonging badge would be the Google News badge. Google issues readers digital badges at the Google News homepage to help visitors learn about their reading habits and to convey shared interests through social networking. Google may also use badges as a reward system to keep users coming back to the News homepage. The Google News badge is designed with multiple levels and is earned by reading articles on certain subjects at a rate higher than other readers. Digital badges can be found in a number of on-line sites and appear frequently in social networking applications. One example is Foursquare, a website that permits users to provide and share location-based data using mobile devices. Within the Foursquare community, badges can be earned for a number of social interactions including the beginner badges for joining and checking in. While not entirely useful for educational purposes, the Foursquare badges may affirm ones achievements and promote group identification.

IV. THE OPEN BADGE INFRASTRUCTURE

The Mozilla Foundation (2012) has proposed a framework called the open badges infrastructure (OBI) for the digital badges ecosystem, which includes: 1) digital badges, 2) assessments, 3) and the open badges cyber infrastructure. In this proposed framework, the digital badge represents skills and abilities of learners and may come in many different forms. Badges may be awarded for a narrow band of achievement and skills or on the other hand may be more comprehensive and awarded for mastery of a set of skills. For example, a badge could be awarded for learning how to use variables in a program language or a badge could be awarded for mastery of the entire language. The scope of the digital badge is established by the issuing organization, which can include educational institutions, training centers, or even employers. Assessment is critical to the badges framework to ensure that the badges truly represent the knowledge and skills badge earners acquired and to convey the information stakeholders. However, the level of rigor for each badge is flexible and it is expected that community-assessment methodology will be utilized. Again the badge issuer sets the assessment rigor for their badges. The third component of the badges framework is the cyber infrastructure. The digital badge infrastructure provides the electronic means to store and retrieve the badges and related meta-data about the badge. The Mozilla open badges infrastructure provides the electronic means to issue, store, display, and endorse on-line digital badges. For example, for institutions that want to issue a badge (issuer) the open-badges cyber infrastructure provides a JavaScript application interface (API) to send an assertion (information about the badge recipient) to the individuals backpack. The backpack is the badges repository and is currently hosted by the Mozilla Foundation. The badge recipient is able to access their badge collection through the Mozilla backpack and push those badges to other networks like Facebook.

In summary, the open badges infrastructure allows organizations to develop and issue their own digital badges. Badge recipients can collect digital badges and then display them on social networks or job sites. The badges represents knowledge and skills obtained outside of the classroom, provides a way for badge recipients to be recognized, and unlocks educational and occupational opportunities.

A. Digital Badges for Learning

Digital badges have been popularized in the gaming industry and through social media systems as symbols of inclusion and status. However, the current generation of digital badges focuses on awards for participation rather than learning and achievement. The transformation of digital badges from simple awards to representations of skills and abilities is the centerpiece of the National 4-H digital badges for learning initiative. As a major leader in youth development 4-H is well positioned to adapt digital badging as a part of the overall recognition model.

According the National 4-H Recognition Model (USDA, n.d.), the systematic recognition of learning provides youth positive reinforcement and the necessary motivation to continue to participate in such learning endeavors and to development life skills [6] [7]. Moreover, the existing 4-H model includes five types of recognition:

- Participation in educational experiences
- Recognition of progress toward personal goals
- Recognition of the achievement of generally recognized standards of excellence
- Recognition through peer competition
- Recognition for cooperation

 ISBN: 978-1-61208-280-6

The 4-H recognition model is adaptive to meet the individual needs of youth and supports a balanced approach to encourage recognition from each of the five recognition categories. The model is also designed to satisfy both intrinsic (internal) and extrinsic (external) motivational needs of individual youth [8]. In practice, 4-H awards ribbons, medals, conference trips, scholarships and many other incentives depending on the state and county of the youth participant and promotes the development of internally relevant skills and knowledge through program participation [9].

With the expansion of social media and the use of digital badging for learning the 4-H organization has the opportunity to expand the current recognition model and reach a wider audience [10] [11]. Digital badges also fit in each of the five recognition categories and may provide individualized motivation for youth. In addition, digital badges permit sharing with potential employers and post-secondary institutions to showcase competencies obtained and awarded through the 4-H experience.

Perhaps the ultimate outward sign of learning today is the diploma. Whether from high school or college the diploma is an important signal that an individual has met some benchmarks of learning. However, a diploma may not provide a record of skills and abilities obtained in the formal classroom. Digital badges, on the other hand, may provide an ideal way for lifelong learners to enhance their learning credentials beyond the realm of formal education [12].

A digital badge for learning can be thought of as a visual representation of accomplishments, certified skills and abilities [13]. The relative advantage of digital badges is that they may provide a more detailed view of what the badge recipient has learned when compared to traditional diplomas and can signify learning in informal environments [12]. Under a digital badging system a young person could display dozens of badges providing a detailed picture of acquired competencies and the skills developed in school and out. Moreover, digital badges may be used to show potential employers not just an earned degree but also a detailed list of demonstrated competences. Such as 21st century workplace skills and important life skills like teamwork, innovation, and leadership.

Digital badges are of great interest to the education community because of their potential to motivate youth to pursue the development of skills and knowledge [11][12]. Badges provide an ideal environment for goal setting whereby learners are challenged to meet the criteria established in the awarding mechanism of the badge. Digital badges also provide a blueprint of educational offerings within a community for those new to a community [11]. In the Boy Scouts for example, badges provide motivation to the earners but just as important it is a recruitment tool for potential badge earners to understand what experiences scouting has to offer. According to Halavais (2012), the process of awarding digital educational badges should lead others to engage in learning.

B. Application of Digital Badges in 4-H

One challenging aspect of the current OBI model is the existence of a large spectrum of badge types that can be earned from very simple check-in and get a badge to more rigorous badges that require comprehensive assessment. To provide outward legitimacy that the 4-H badges represent robotics and engineering based skills and abilities the devised 4-H badges assessment mechanism is moderately robust. In general, the steps for youth participants to earn a badge are: 1) identify an adult facilitator who will conduct the assessment, 2) complete the curriculum, 3) complete an engineering notebook, and 4) complete and submit an on-line survey and provide the adult facilitator the engineering notebook for review.

Following the OBI model the current 4-H Digital Badge ecosystem is comprised of five badges all focused on robotics including: 1) Robot Hands, 2) Robot Movement, 3) Mechatronics, 4) Robot Platforms, and 5) Robot Competitions. The badges can be earned by youth aged 9 to 15 that participate in one of the five robotic programs.

To earn a robotic badge youth must provide evidence of progress in four main areas (science abilities, workforce skills, science knowledge, and engineering performance). Science abilities include skills like observation of a phenomenon, prediction, and redesign. Secondly, youth will display improvements in 21st Century Workforce Skills including critical thinking and problem solving skills. Third, youth will provide evidence of learning gains in science knowledge and big ideas related to the projects including friction, balance, circuits, and electricity. Finally, youth will be assessed on the completion of program benchmarks like construction and functioning of the robotic system or sub components. Performance in all four areas is measured by a 22-item 4-point Likert-type scale, three to four essay questions and a review of engineering notebooks that youth complete in the project, and the approval by adult facilitators that work with youth in the five badge areas see "Table 1." for a list of criteria and assessment procedure.

TABLE I. ASSESSMENT FOR BADGE ISSUANCE

Badge Requirements		
Evidence	***Example of Skills***	***Measurement***
Science, Engineering, and Technology Abililites	build/construct, communicate results	16 self-reflective Likert-type survey questions
21st Century Workforce Skills	critical thinking, decision making	6 self-reflective Likert-type survey questions
Science Knowledge	scientific habits of the mind	engineering notebook entries
Performance Benchmarks	build a robot hand, program a robot to move forward	3 to 4 essay type self-reflective questions

The survey instrument was field tested with 30 youth ranging in age from 9 to 15 in two US states. In addition, 15 adult facilitators also took part in the piloting of the instrument. The goals of the piloting process were twofold. Goal one was to introduce the digital badges issuance and

assessment framework to youth that had recently completed one of the five robotic projects. The second goal of the pilot was to test each survey question with respondents thereby reducing the overall measurement error. To test each question on the survey youth were given instructions to answer the questions by circling the corresponding number in the 4-point Likert-scale. Next, the researchers read each question with the youth and asked their level of understanding and what could be done to improve the questions. This was done for all 22-questions and the four open-ended questions. Results of the field test showed that 12 of the survey questions needed to be reworded with simpler words like substituting *autonomous* with *runs by itself* or additional examples were needed like (*why and how things work*) was added to the questions *I can apply basic scientific principles to my 4-H robotics project*. In addition, the Likert-scale headings were changed for four of the survey items. In addition, a focus group interview garnered important feedback on the proposed process to earn and issue badges from the youth and adults point of view. Most importantly the adult facilitators felt that they could conduct the required assessments to enable the issuance of digital badges to youth that they teach. Upon the completion of the pilot process the surveys were incorporated into the 4-H digital badges cyberinfrastructure and will be tested with additional groups.

V. CONCLUSION AND FUTURE WORK

While the OBI is still at the very beginning of implementation in industry and educational institutions, it has the potential to bridge formal and informal learning environments by recognizing skills and competencies through learning ecosystems. Youth serving programs like 4-H will implement badges for their members with the expectation that it will enable youth to compete for employment and academic pursuits. More research is needed to determine the role of badges on motivation and goal setting. In addition, developers of the ecosystem will have to establish validity and reliability of individual badges through assessment processes.

REFERENCES

[1] K. Carey, "Show me your badge," The New York Times, Nov. 2012, avaialable from http://www.nytimes.com/2012/11/04/education/edlife/show-me-your-badge.html?pagewanted=all&_r=0 , <retrieved 03, 2013>.

[2] US Department of Agriculture. "National 4-H recognition model," available from http://www.national4hheadquarters.gov/library/4h_recmo.pdf, <retrieved December 2012>.

[3] National 4-H Council, "4-H History," available from http://www.4-h.org/about/4-h-history/, <retrieved December 2012>.

[4] National Research Council, "Learning Science in informal environments: People, places, and pursuits," Washington, DC: The National Academies Press, pp. 11-17, 2009.

[5] E. Goligoski, "Motivating the learner: Mozilla's open badges program," Access to Knowledge, vol. 4(1), 2012, pp. 1-8.

[6] J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman, "I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application," Proc of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11). ACM, New York, NY, USA, 2011, pp. 2409-2418, doi:10.1145/1978942.1979295 http://doi.acm.org/10.1145/1978942.1979295.

[7] B. Boyd, D.Herring, and G. Briers, "Developing life skills in youth," Journal of Extension, vol. 30(4), available from http://www.joe.org/joe/1992winter/a4.html, 1992, <retrieved December 2012>.

[8] M. Norman and J. Jordan, "Targeting life skills model," University of Florida IAFS Extension, available from http://www.csrees.usda.gov/nea/family/res/pdfs/Targeting_Life_Skills.pdf , <retrieved April 2013>.

[9] E.A. Quarrick, and R.E. Rankin, "Intrinsic motivation in 4-H," Journal of Extension vol. 3(1), available from http://www.joe.org/joe/1965spring/1965-1-a6.pdf, 1965, pp. 42-50, <retrieved December 2012>.

[10] S. Forbes, "The 4-H incentive system," Journal of Extension, vol. 30(3) Article 3RIB1, available from http://www.joe.org/joe/1992fall/rb1.php, 1992, <retrieved December 2012>.

[11] J. Antin, and E. Churchill, "Badges in social media: a social psychological perspective," in proceedings of the 2011 annual conference on Human factors in computing systems (CHI 2011)," Vancouver, BC, Canada, May, 2011, pp.1-4.

[12] A. Halavais, "A genealogy of badges: Inherited meaning and monstrous moral hybrids," Information, Communication, & Society, vol. 15(3), 2012, pp. 354-373.

[13] J. Selingo, "Colleges need some big ideas to drive change from within," Chronicle of Higher Education, vol. 58(7), September 4, 2011, pp. A1-A2.

[14] J. Young, "Badges earned online pose challenge to traditional college diplomas," Chronicle of Higher Education, vol. 58(19), May 11, 2011, pp. A1-A4.

 ISBN: 978-1-61208-280-6

A Mobile Learning App for Driving Lessons

J. Molina-Gil, C. Caballero-Gil, P. Caballero-Gil
Department of Statistics, Operations Research and Computing
University of La Laguna
Tenerife, Spain
Email: {jmmolina, ccabgil, pcaballe}@ull.es

A. Quesada-Arencibia
Institute of Cybernetic Science and Technology
University of Las Palmas de G.C.
Las Palmas, Spain
Email: aquesada@dis.ulpgc.es

***Abstract*—Mobile learning, or m-learning, is a new step in e-learning that offers novel opportunities for learning beyond traditional classrooms and computer-based tools, thanks to the use of mobile devices. The mobility of digital technologies may be applied to change the nature of the relationships among teachers, students and learning objects. In this paper, we propose a mobile application, called Planning2Drive, as an m-learning example that shows the potential of smartphones in the learning process, mainly because they can be used anywhere, any time.**

***Keywords*-Mobile learning; m-learning; education; mobile devices; outdoor learning.**

I. INTRODUCTION

Mobile learning can be defined as the use of wireless mobile technology for e-learning [1]. M-learning is a subset of e-learning through mobile computational devices. It allows the access to a learning environment from anywhere, at any time. Therefore, learners can use wireless mobile technology both for formal and informal learning, so that they can get additional and personalized learning materials from the Internet and/or from the host organization.

Most research on e-learning and m-learning is focused on theoretical learning. In particular, in the literature we can find different proposals on how to generate contents, exercises, activities, etc., for e/m-learning platforms. However, in general the practical part of the lessons is what provides a richer education and promotes lifelong learning.

A practice environment is defined as a situation where students apply their knowledge to practice, learn key skills and achieve the required competencies. Thus, it can be seen as a way to reinforce learning. Despite the great progress in mobile devices and their growing influence in our society, they have not yet been proposed as a learning tool in a practice environment. In this paper, we show an example of how e-learning based on mobile technology can be used in the teaching/learning process, through a learning case study and the analysis of its educational potential.

Practical lessons, like those followed by students who are preparing for the driving license test, do not allow students to keep track of their progress and subsequently to fix the points where mistakes were made. These aspects are very important in any teaching process, not only to ensure that teachers have reached the learning goals but also to give students the necessary trust to face the practical driving test.

In this paper, we present a mobile application called Planning2drive, which allows students to create learning sequences from their own experience as well as from the output generated by other students. It creates a collaborative environment where not only students learn from their own mistakes, but also from the mistakes of other students. In addition, students can track their progress as well as compare them with other students who are preparing or have taken the test. This work is focused on the practical driving test in UK for being a fairly easy system due to its properly regulated assessment model.

This paper is organized as follows. Section II describes related research about e-learning and m-learning. Requirements and explanation of the proposed application are included in Section III. In Section IV the potential of this new technology as tool for teaching and learning is analysed. Finally, Section V includes conclusions and future work.

II. BACKGROUND

Nowadays, information technologies, such as email and instant messaging, are widely used to facilitate communication and collaboration. In [2], the authors state that technology can be effective in the learning process only when it meets specific educational needs. This happens when the activities that learners perform are active, constructive, intentional, authentic, and cooperative. In addition, access to multimedia technology has previously been shown in [3] to improve cognitive engagement and cognitive absorption in users. On the other hand, Ellison et al. [4] states that instructors should investigate which are the software implementations that best support the pedagogical goals and needs of students in advance.

Many papers study the influence of m-learning in the teaching-learning process. For instance Hardless et al. [5] presents a research project on training, education and sharing of experiences among mobile people in a professional environment. That work shows the need for new forms of education in which students can participate where and when they choose, and evidences how third generation cellular can provide them. In [6] the author presents a study of

 ISBN: 978-1-61208-280-6

the mobile media revolution on instructional design and learning effectiveness in nursing education. In [7], Alexander proposes the use of iPhones to take attendance. In particular, according to that proposal, as students enter the classroom, instead of writing their name on a sheet, they simply enter their ID number and a specific class number into an iPhone application. In order to prevent students from logging in from home or outside class, the application uses GPS location data and checks which router the students have logged in to.

Finally, Traxler [8] presents a preliminary attempt to address the issue of definition and conceptualisation of m-learning, and draws on recent research examining case studies from the UK and elsewhere. Such a research shows how mobile learning can transform the delivery of education and training.

III. Research approach

A. Understanding the Official DSA car practical test

The Driving Standards Agency (DSA) is an executive agency within the UK Department for Transport. Its mission and primary aim is to promote road safety by influencing driver and rider behaviour. It delivers tests from around 400 driving test centres and 140 theory test centres. The DSA provides the possibility that students book the practical test online, so that they can check whether their preferred date and time is available and book it by phone or through email. On average, those who pass the test have taken around 45 professional lessons and 20 hours of private practice.

When students get to the test, they are asked if they want to be accompanied by their driving instructor or another person during the test, and after it for the result and feedback. The test lasts from 28 to 40 minutes. During it, students drive in various road and traffic conditions and are asked to drive independently for approximately ten minutes, by following either traffic signs or a series of verbal directions, or a combination of both. They can also be asked to complete certain test manoeuvres such as turning in the road, parallel parking, reversing around a corner, or doing a controlled stop. If students commit more than 15 driving faults or a serious or dangerous fault at any time, they fail the test.

B. Planing2Drive Architecture

Planning2drive is a mobile phone application whose goal is to monitor students during the teaching-learning process and the preparation for the driving test. A Beta version is already available for both iPhone and Android platforms. In addition, the application allows to maintain student-teacher communication. In the teaching-learning process two main roles are distinguished within the application's functionality:

- **The teacher**, who monitors and supports students throughout the learning process. He/She is responsible for signalling mistakes during practices, using the Planning2Drive platform.
- **The student**, who can receive professional and private practices. Besides, he/she can check all the mistakes committed in the practices that have been marked by the teachers using the website or a mobile phone.

Planning2drive app allows students to be best prepared for the driving test so that they can find it less difficult. Moreover, through its use, students can reduce the number of lessons, and consequently save money.

In this section, we outline several use cases of the proposed application, suited for teachers and students.

1) Planing2Drive Website: The website [9] allows to create a virtual classroom where all the professionals of driving training can be registered. This will be the meeting point between teachers and students. In order to ensure the e-learning platform, in addition to registering, the professionals have to prove their ability to teach driving practices. Once the website administrator has checked the data, the website shows the contact details of the teachers to the students, so that they can book lessons with them. Each teacher has a calendar that is administered by him/herself, indicating the available hours. Students registered in the system can select the teacher they want and ask for a lesson using the timetable set by the teacher.

In addition, each teacher shows on the website a set of practices performed with his/her students, including the committed mistakes. The teacher is free to publish all, some, or none of the practices, but the recommendation is that the teacher only publish the practices that can help students in the learning process. In any case, student identifications are never shown with the published practices.

On the other hand, teachers also publish their students' exam practices, both the passed ones and the failed ones. In this way, students can see the most common faults in practice exams and learn from them. In addition, it will help them not to get nervous if they make a mistake because they would have seen how other students have passed the exam even with some errors.

Students who register in the system, in addition to book lessons with teachers, can keep track of all the practices they have done with each teacher. In this way, they can strengthen their learning, not only when they make a mistake and the teacher indicates it, but also when they achieve observational learning through reviewing their lessons and mistakes.

2) Planing2Drive App for Professional Lessons and Exams: Once the student/teacher relationship is set, the teacher can begin the process of tracking practices of each student. The application is very intuitive and easy to use so, the teachers have no problem with it, and the corresponding learning curve indicates a quick progress in learning.

When the teacher logs in at the application, he/she is shown a list of students who have booked a practice with him/her that day (see Figure 1A). As shown in Figure 1B, once the teacher selects a student, he/she can see the student's contact information, like phone number and email.

 ISBN: 978-1-61208-280-6

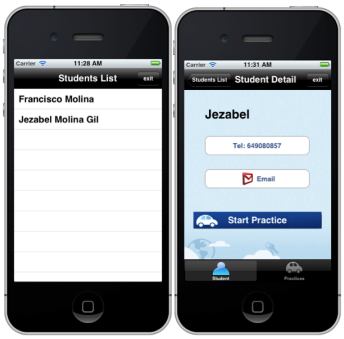


Figure 1. Teacher Interface: Students List and Students Details

	Junction: observation, speed
	Respect the stop
	Improper turning right
	Respect the give-way
	Respect pederatrian crossings
	Improper overtaking
	Incorrect following distance
	Signals necessary
	Respect cycling lane
	Jump red light

Figure 2. Common Error Table

Figure 3. Errors in a Practice

The first step for a practice lesson is that the student and the teacher agree a meeting place, because there is no school or classroom. The proposed application facilitates contact between the student and the teacher in case of unexpected change of plans, such as delay, cancellation, etc. It also provides an interface that allows the teacher to start a practice and mark possible faults committed by the students. In order to mark the mistakes during the practices, the DSA Driving Test Report provides a list containing the most common mistakes, but teachers are also permitted to create and introduce new mistakes. When the practice lesson starts, the teacher pushes the correspondent button, and in that moment the app shows the map with the current location and a list of possible errors nearby. Each time a student commits a mistake, the teacher pushes the icon corresponding to the mistake and that mistake is shown on the map. Figure 2 summarizes some mistakes that the application contains and the corresponding icons that represent them. Locating the mistakes on the map is beneficial for the student because in this way he/she can pay more attention to avoid repeating the same mistakes in the same places, especially when the student is carrying out a test in the area. The application contains a field that allows distinguishing between the two types of practices: training and test. This not only allows improving the student's knowledge but also allows sharing with other students' information about the possible mistakes during the exams and about the usual exam routes.

3) Planing2Drive App for Private Practices: The best way to learn is through the combination of professional lessons and private practices. Currently, the students try to get as much experience on the road as they can. One of the problems they face is that they only receive feedback during the lessons or when they finish them and speak with the teacher. The proposed application provides a tool that not only allows obtaining information about the mistakes the students commit, but also reviewing such feedback from anywhere, including from home. Students can get information about the place where the mistakes were committed and afterwards repeat the same route as many times as they want either in professional lessons or private practices.

Each student who has installed the application can use it for two purposes. On the one hand, the student can record the private practices following the same process that a professional teacher. In this case, the person responsible for overseeing these private lessons, is responsible for following the teacher's role specified in section III-B2. On the other hand, the student has access to data about all types of practices (professional or private) that have been recorded, so therefore he/she can check the committed mistakes during them. Figure 3 shows an example of a recorded practice that a student can visualize. In that example the student has committed three faults, two of them involving driving faults while the other is a serious fault. The first corresponds to a manoeuvre that had not been properly signalled. If the teacher used the tool correctly, the mistakes appear in the exact place where they were committed. As can be seen in the image, the second fault occurred while the student was driving on a highway, where the student exceeded the limit speed, which in this case was set at 70km/h. This would mean a failing in the test. Finally, there is a signal indicating that the student has not respected the safety distance with the vehicle ahead.

Besides, the students have a section in the website where they can share experiences in their practices. The most important aspect is the possibility to establish cooperation

 ISBN: 978-1-61208-280-6

among students. Students can explain their experience with the tool during the teaching/learning process, and they can also share with other students which teachers they have liked most and why. They may also indicate which practices have been more useful for them. It provides a cooperative environment where students can share their experiences.

IV. Pedagogical Value

The purpose of this study is to investigate the pedagogical value of the Planning2Drive application by identifying its relationship with factors that may have the potential for improving learner outcomes. For the purpose of this study, the pedagogical value has been defined on constructivist principles in the learning process, by retaining and using information about committed mistakes.

An important aspect for the learning process is that students have opportunities to receive feedback, especially to correct faults, and to incorporate that information into their further understanding and/or performance. Constructive criticism, performed effectively, is a productive educational activity. However, in the proposed application, the public nature of feedback, particularly criticism, can complicate its effectiveness as a form of pedagogy. Thus, possible defensiveness and embarrassment of students have to be avoided by providing complete anonymity in the data offered to other students. Planning2Drive reaches this goal by showing practices and exam examples without showing the students' names. On the other hand, the students can review their own practices, in a private way.

Chickering et al. [10] identifies seven principles that are types of teaching and learning activities needed to improve learning outcomes. Some of those principles are met in Planning2Driver. First, in order to state a good teaching procedure, it promotes reciprocity and cooperation among students. This is accomplished through the web platform that offers students the opportunity to share their experiences. Another proposed principle is encouragement of active learning. In the proposed tool, students are able to see the immediate results of their professional practices and create their own practices. Students can also ask teachers to perform some of the practices that other students have categorized as difficult in order to analyze and compare their own results. Furthermore, as recommended, a prompt feedback is give throught he proposed application. In particular, during the practices, the teacher indicates what the student has done well and wrong so that after the student finishes the practice, he/she can immediately visualize the commited mistakes in his/her mobile phone and review them with or without the teacher. One of the most important aspects of the used technology is that it allows to respect diverse ways of learning. Students can fix their pace of learning by doing less or more professional practices depending on their own needs.

V. Conclusion and future research

Planning2Drive is an m-learning tool that allows tracking driving lessons. This study has demonstrated that the proposed tool presents an important educational content and shows once again the importance and potential of mobile technology in the teaching/learning process. Planning2Drive is a tool yet under development. In particular, a Beta version is already available which will be put into operation soon in UK, so wide data and feedback about its usefulness will be obtained from teachers and students. The anaysis of such results and consequent improvements of the tool are part of the future research.

Acknowledgment

Research supported by the Ministerio de Economa y Competitividad and the FEDER Fund under Projects TIN2011-25452 and IPT-2012-0585-370000, and the FPI scholarship BES-2009-016774, BES-2012-051817, and ACIISI-BOC Number 60.

References

[1] M. Sung, et al. "Mobile-IT Education (MIT. EDU):m-learning applications for classroom settings". Journal of Computer Assisted Learning Vol. 21, N. 3, 2005, pp. 229237.

[2] D. H. Jonassen, J. Howland, R. M. Marra, and D. P. Chrismond, "Meaningful Learning With Technology". Prentice-Hall, 2008.

[3] R. Agarwal and B. Karahanna, "Time Flies When You Are Having Fun: Cognitive Absorption and Beliefs About Information Technology Usage". MIS Quarterly, Vol. 24, No. 4, 2000, pp. 665-694.

[4] N. B. Ellison and Y. Wu, "Blogging In The Classroom: A Preliminary Exploration Of Student Attitudes And Impact On Comprehension". Journal of Educational Multimedia and Hypermedia, Vol. 17, No. 1, 2008, pp. 99-122.

[5] C. Hardless, J. Lundin, and U. Nuldn, "Mobile competence development for nomads". System Sciences, Proceedings of the 34th Annual Hawaii International Conference on Source: IEEE Xplore, DOI: 10.1109 /HICSS.2001.926229, 2001.

[6] M. Maagi, "iPod, uPod? An emerging mobile learning tool in nursing education and students satisfaction". Australasian Society for Computers in Learning in Tertiary Education (ASCILITE), Sydney, Australia, 2006.

[7] B. Alexander, "Using smartphones to track attendance". Liberal Education Tomorrow, 2010.

[8] J. Traxler, "Defining Mobile Learning". Proceedings IADIS International Conference Mobile Learning 2005, Malta, 2005, pp. 261-266.

[9] http://www.planningtodrive.com/, 2013.

[10] A. W. Chickering and Z. F. Gamson, "Seven Principles For Good Practice In Undergraduate Education". The Teaching, Learning and Technology Group (TLT). Retrieved on September 6 from http://www.tltgroup.org/Seven/Home.htm, 2008.

 ISBN: 978-1-61208-280-6

Monitoring Activities in an E-Learning 2.0 Environment

A multi-agents system

Henda Belaid-Ajroud
LIPAH
Faculté des Sciences de Tunis,
Le Belvédère, Tunisie
henda.ajroud@fst.rnu.tn

Bénédicte Talon, Insaf Tnazefti-Kerkeni
LISIC
University of Littoral-Côte d'Opale (ULCO)
Calais, France
{Benedicte.Talon/Insaf.Kerkeni}@univ-littoral.fr

***Abstract*—This paper deals with the monitoring of students and teacher's activities in a collaborative pedagogical environment. The specificity is that the learning platform is an E-Learning 2.0 environment. It is complex to track activities because the E-Learning 2.0 environments are adaptive and the tracking is not easy to be anticipated. This paper first presents the general context. Then, it describes the specific pedagogical method and environment and the studio used to generate the pedagogical environment. It continues with the architecture proposed to track activities. Then, the collected traces and provided indicators are presented.**

Keywords-Collaborative environments; E-Learning 2.0; Web 2.0; Monitoring of activities, Multi-agents Systems

I. INTRODUCTION

There are various forms of collaborative learning platforms. Among them, environments reflected in the literature under the name E-Learning 2.0 [1], allow teachers to exploit Web 2.0 applications to construct educational environments. They are a combination of specific features of Web 2.0 applications: Forum, Wiki, documents management, Blog, etc. The advantage is that Web 2.0 tools are directly available, generally free and they can easily be used in an educational setting [2][3]. The empowerment of teachers is at the heart of these solutions because the Web 2.0 tools runs on all browsers. The actors may be released from the administration constraints generally associated with traditional architectures and easily share access to their resources.

Researchers of the University of Littoral Côte d'Opale (ULCO) and the University of Picardie Jules Verne (UPJV) work on a Pedagogical Engineering Studio (PES) called MACADDAM (MAui for Computer Aided pedagogical Design bAsed on MAETIC) [4]. MAui is an acronym for "Méthode de conception de dispositifs pédAgogique Utilisant l'ethnographIe" derived from [5] and MAETIC for "Méthode pédAgogique instrumEntée par les Technologies d'Information et de Communication" [6]. This PES assists teachers in their efforts to design educational systems. Environments generated by MACADDAM are dedicated to education through collaborative projects.

However, in the current version, services offered by the generated pedagogical environments have a lack of monitoring. However, in the field of collaborative learning, management of traces is important because it is necessary to analyze information about actors and their activities [7]. It provides the trainer with accurate and adequate information to track individual and collective participation.

Tracking systems collect traces and interpret them with the computation of indicators. In fact, the indicators enable the evaluation of the learner. This evaluation can be individual or collective which is the evaluation of the learner's group. Different types of indicators can be defined such as indicators to know acquired knowledge during the learning activity or to know the communication and the interaction between the learners of the same group.

Tracking systems have already been developed such as APLUSIX [8], ACOLAD [9], aLF [10], SPLACH [11], TrAVis [12] and other ones. However, each tracking system has a different target. All of these tracking systems are platforms-specific and none of them can be used with a Web 2.0 learning platform. This is the reason that led us to develop a system dedicated to tacking activities in an E-learning 2.0 environment. This developed system is a multi-agents system (MAS) [13] and is coupled to the MACADDAM studio and implements functionalities to keep detailed history of students' actions, group of students' actions and teachers' actions performed on the E-Learning 2.0 platform.

Section II describes MAETIC, the MACADDAM studio, specifies the requirements, and explains the choice of a multi-agent system coupled to this studio to track activities before discussing some related works.

Section III describes this collaborative learning-oriented agent system. Section IV deals with collected traces and provided indicators. The last section concludes and presents some perspectives.

II. THE MACCADAM STUDIO

In this paragraph, we describe MAETIC method and the specificities of the MAETIC environments.

A. MAETIC and its pedagogical environments

Teachers of ULCO and UPJV have designed educational environments using Web 2.0 tools. A study of these environments allowed to extract a pedagogical method called MAETIC. MAETIC is dedicated to the management of project-based pedagogy in-group. It was validated through successive evaluations [14]. MAETIC aims at developing professional skills (transverse and domain skills) and guides groups of students in all stages of the project. The part of the

 ISBN: 978-1-61208-280-6

system dedicated to students is called "MAETIC e-suitcase". The concept of e-suitcase refers to an environment "which is not bound to a fixed place of education." The e-suitcase includes access to the teacher's logbook (important information, activities of the session, etc.), access to teacher's resources (course materials, exercises to do, etc.) and access to student's logbook. The teacher is informed on the progress of the project via the group's logbook. The logbook describes the life of the project. The developed deliverables, the report of the activities and information on the project are available to the teachers and other members. The logbook is managed regularly. The method is described in [6].

The part of the pedagogical environment dedicated to the teacher is called "MAETIC Toolbox". A toolbox provides mechanisms to fill the teacher's logbook, to check students' logbooks, to comment on their work, to assess their work, etc.

So, in a MAETIC environment, the teacher:

- Uploads articles and resources on his/her logbook. This logbook informs students about the life of the teaching unit.
- Posts general comments about the work and about its progress.
- Handles tools that enable him/her to communicate with students,
- Oversees the work of the groups. Thus, he/she can view, download and comment the activities of the groups via the groups' logbook.

The student

- Consults or download resources. They are available and accessible via the teacher's logbook. Each student must consult the teacher's logbook before each session.
- Uses tools allowing him to communicate with other students or with the teacher.
- Realizes activities related to the planned project. The teacher helps to define these activities. The report and implementation of these activities are recorded on the group's logbook.

Fig. 1 synthetizes the different interactions.

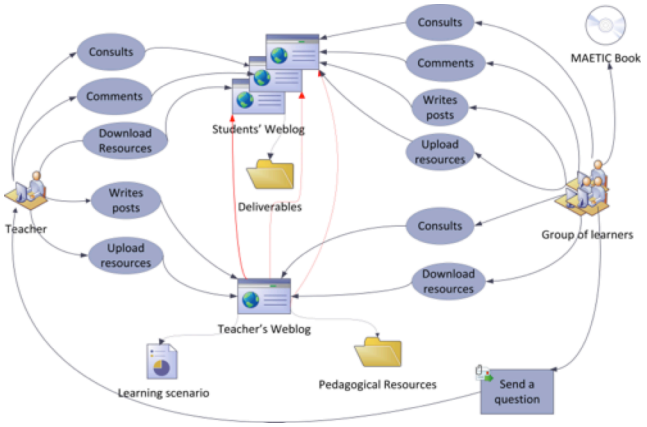


Figure 1. Users interactions in MAETIC

The next paragraph describes the aim and structure of the MACCADAM studio.

B. *MACCADAM*

MACADDAM is a Pedagogical Engineering Studio (PES). It helps teachers to deploy their own educational environments dedicated to the implementation of MAETIC. The studio assists them in the formulation of needs (design support) and releases them of repetitive and tedious tasks associated with the deployment of a teaching environment (development support). The environments are designed and are instrumented using Web 2.0 tools [4].

The PES assists the teacher in the formation of the pedagogical script and generates the e-suitcases and the toolbox.

An assistance module drives the teacher throughout the design process. A generator module generates either a generic environment or a customized one (according to the teacher profile and the target skills). The studio offers the ability to select and identify activities to be implemented in the script. However, resources and activities appropriated to the training area are under the teacher's responsibility.

The next paragraph describes the needs in term of activities tracking and justifies the choice of a multi-agent architecture.

C. *PES needs and contribution of multi-agent systems*

Communication between teachers and students in a MAETIC environment is mainly done via Weblogs. Weblog technology was chosen for its usability. The weblog is accessible to everyone. In addition, weblogs can create a social relationship between students and seem to facilitate the students' writing through the "posts" [15]. However, Weblog technology provides a very consistent material, easy to collect but more difficult to analyze. The time spent by the teacher to monitor and analyze the activities of the student is higher than the time spent in the traditional classroom [16].

The statistics about interaction enable the evaluation of the group's life and its evolution. We are particularly interested in the status indicators of progress and sustainability of the group.

Among indicators, one can cite:

- Identification of work overload for a given student so he/she can be exempted to perform some activities,
- Possibility of extending or shorting completion deadlines of an activity (change the training schedule),
- Assessment of the state of completion of an activity or a task,
- Evaluation of social relationships and productivity of a student, and so on.

Our main goal is to provide a relevant assistance for the progress of the project activity. On one hand, we want to help students in the realization of the project in collaborative learning, and on the other hand, we also want to assist the teacher in the monitoring of both individual and collective activities.

 ISBN: 978-1-61208-280-6

Our goal is to develop a system that is able to collect and analyze data from the project activities. The system must be able to analyze the use of the environment and the data generated in this environment (forum, mail, meetings, etc.).

To design the observation and assistance system, we have opted for an incremental and iterative approach. The environments generated by the MACADDAM studio will be equipped with this system. It is based on a multi-agent architecture described in the next section.

The choice of a multi-agent architecture for the observation and assistance system is motivated by several reasons:

- From a programming point of view, it is possible to add new agents or modify the behavior of existing agents without affecting the overall structure. In a research context, the possibility of change is a considerable advantage because it allows an iterative and incremental development.
- We are interested in providing the ability to solve distributed problems in a multi-agent architecture. To meet the specifications outlined in the previous section, we propose to identify agents that are specialized in observation tasks and others that are specialized in assistance tasks.

Moreover, in our case, we are faced with a distributed environment. The multi-agent approach allows to have distributed agents being able to communicate.

D. Related Works

Before discussing the proposed architecture, we review the approaches presented in the literature that provide multi-agents platforms for collaborative learning. In the field of Artificial Intelligence and education, several approaches have been developed.

For example, Guizzardi and al. [17] have developed a Peer-to-peer system called "Help & Learn". This system was modeled using an agent-oriented language called AORML [18]. It is an open system that is designed to support the extra-class interactions between learners and tutors. "Help & Learn" is limited to providing assistance to learners who request it. Other systems have been developed. Fougeres and Ospina [19] have proposed a based-agent mediation system for the project management platform called iPdagogique. This system, modeled in AUML, serves as an interface between the human and the application to enhance their relationship and is used to promote collaboration among users. Recently in [20], the authors presented a model for an adaptive multi-agent system for dynamic routing of the grant's activities from a learning environment. This model allows the assignment of activities taking into account the specialization of learners, their experience and the complexity of activities already taken. None of these three systems cares of monitoring learning and therefore, cannot trace user's activities.

Mbala and al. [21] have developed a multi-agent system called SIGFAD to support users in remote education. SIGFAD is modelled using the MASE methodology and uses the JAM model for building agents. It is interested in monitoring learning.

However, it is not sufficiently independent and does not start up alerts to prevent tutors if there's a problem with a learner or group.

In the next section, the architecture of the tracking system is described.

III. PROPOSED ARCHITECTURE

The system uses tools to perform the following functions:

- Give information on the process, the resources and the learning modalities (databases, catalogues, etc.);
- Communicate and coordinate the actors (forum, chat, email, etc.);
- Adjust the schedule of activities during the training;
- Monitor, guide and control (logbook, support the link between learning, support resource allocation, etc.).

We are interested in providing observation and support tools to ensure the following functions:

- Course construction and management (self-diagnostic tools, course management software, etc.);
- Review and validation (assistance to individual and collective reviewing, etc.).

Thus, we have identified three spaces: a teacher's space, a student's space and a group of students' space. Each space has a descriptive name, functionalities, educational resources, technological tools (Web 2.0 tools) and functional tools for the observation of use.

The MACADDAM studio generates the following tools to create the pedagogical environment:

- Technological tools are based on web 2.0 technologies. They are tools that the actors need to perform activities in their space. These include, for example, the student's logbook and the teacher's toolbox.
- Functional agents are tools for the observation of use. These tools mark out the behavior of students, groups of students and teachers. They analyze traces too.

We present in Fig. 2 an overview of our environment, with the different agents present in the system. We distinguish two types of user, namely the student and the teacher, and three workspaces, namely, the teacher's space, the student's space and the group's space. We associate an agent to every user. The agent is located on the server. This agent migrates on the user's workstation as soon as he/she connects. The agent is coded as a Java program; applets are programs living on the server which run on the client. This technology allows a user to run his/her agent directly from his/her client. The agent superintendent of the space groups lives on the server. It is active as soon as one of the students of the group is connected. This agent is in charge of providing the environment meta-information on the activities (beginning date, end date, concerned persons, used tools, etc.) and on the forums (beginning date, end date, etc.). We were also interested in the supervision of the interactions between the various users during the formation. We have defined an agent overseeing every communication tool (email, forum, chat, blog). This agent supervises all the

 ISBN: 978-1-61208-280-6

actions of a user during the session. Every event is dated and commented.

An agent manages the group's space. It aggregates information about connections, activities and communication. This information allows to appreciate the life of the group, the productivity of the members and the level of realization of the educational activities. An evaluation agent analyses this information to estimate the lifecycle of the formation. This agent can make objective decisions about modifications of the calendar of activities.

The main agents of the system are the following ones:

- A-LEARN: It supervises the student's space. It allows the supervision of all activities of a student and provides an overall evaluation of his behavior during a training session;
- A-TEACHER: It supervises the teacher's space. It supervises educational resources loaded in his/her logbook, access to group's logbooks and used tools to communicate with students.
- A-GROUP: It supervises the group's space during a session. It supervises actors' activities during a session. It indicates the degree of respect, the success rate, the start date and end date of an activity. This agent provides the list of present and absent students and statistics concerning the progression of each activity. It reminds students about deadlines and notifies the late groups by sending alerts.
- A-TOOL: It supervises tools and provides statistics about their use (Email, Chat, Forum, blog, CVS, etc.).
- A-EVAL: This agent aggregates the information collected in order to structure them and to present it to the Evaluation module of MACADDAM studio

IV. TRACES

The management of the interactions taking place within the educational system is done thanks to the collected traces.

Generally, a trace represents the interaction of the user with the system. As it was defined by G. Dyke [122], "informally, the traces of an activity are the marks which that activity leaves on the environment". In this educational platform, we adopt the definition of a trace given by K. Lund and A. Mille in [23]. According them[1], "a trace is a sequence of observations located in time. The observation is either an interaction between humans mediated in various ways by computer, or a sequence of actions and reactions between a human and a computer".

When using the educational platform, traces are collected. The processing of traces (traces observed from various sources - from server-side as well as from client side) provides knowledge about the activity called learning indicators.

A. *Collecting traces*

Each action of the learner or of the teacher can be traced through the agents defined in the system as shown in Fig. 1. The traces are dealing with synchronous or asynchronous interactions of the user with the system.

For example, to facilitate the control of emails between learners, we defined five types of messages:

- Proposition: A proposes something to B (for example, to become a member of the group, to perform a task, etc.)
- Proposition rejection: B uses it when he rejects the received proposition.
- Acceptance of proposition: B uses it when he accepts the received proposition.

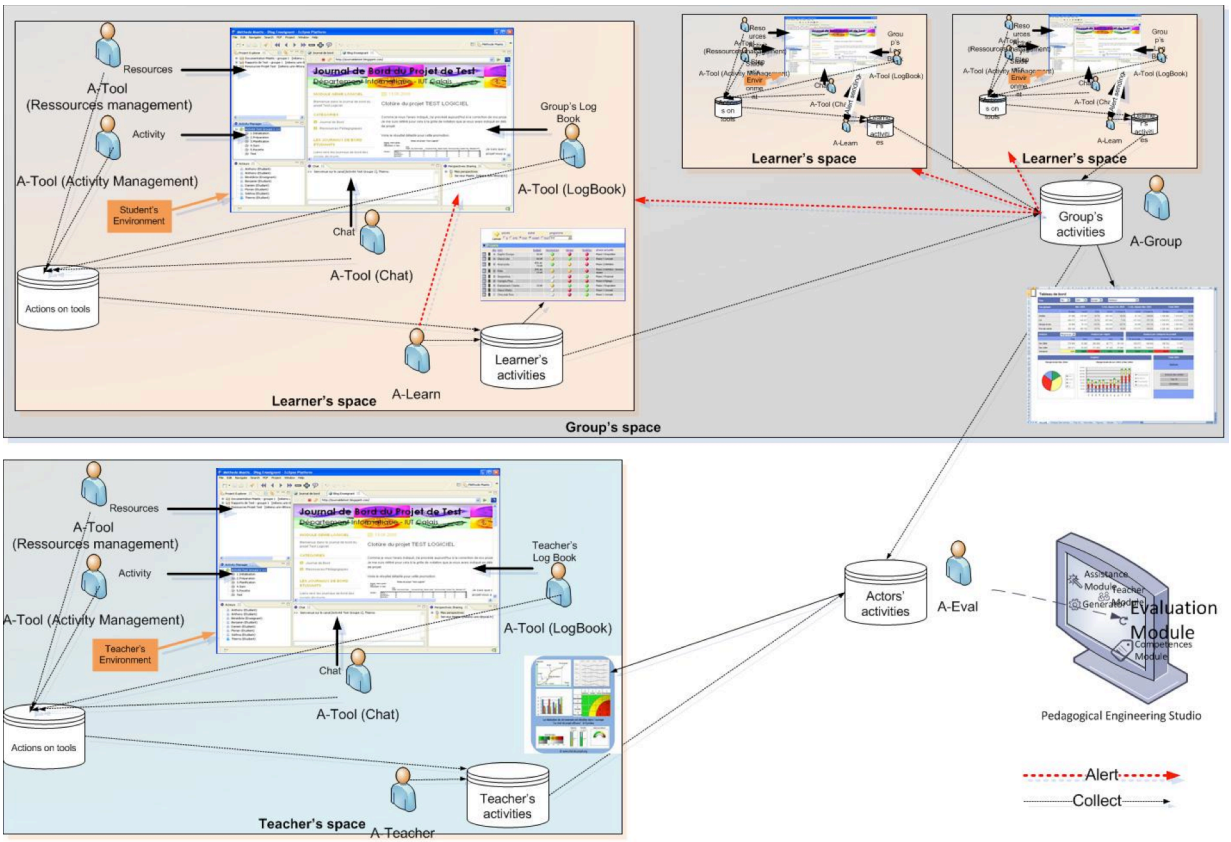


Figure 1. An overview of the environment

 ISBN: 978-1-61208-280-6

- Information: It is used to communicate information or a result. Reply to this message type is optional.
- Help: It is used when A wants some help from B.

Therefore, the trace observed after transmitting an email includes: the subject, the date and time, the sender, the receiver and the content. When observing the blog of a group, the A-TOOL agent notes for each access of the blog: the date and time, the acceded resource, the learner who acceded, the operation done.

B. Providing indicators

Indicators allow to know:

- Who are the people interacting with the system?
- What has been handled?
- When (time or duration).
- How: through which tool?

The two main roles of the indicators in the platform are alert and appreciation. Table 1 presents some indicators.

TABLE I. SOME INDICATORS

Indicator	Purpose	Target user
Percentage of activities carried out per learner	Alert	Learner/ Group of learners
Sleeping learner	Alert	Learner/ Group of learners
Duration of the consultation of each resource per learner	Appreciation	Teacher
Duration of the realization of each activity per learner	Appreciation	Teacher

C. Use of the defined indicators

The alert indicators are used by A-LEARN agent and by A-GROUP agent. They inform the learner and the group of a problem. Communication between members allows to see what is happening and to find a solution. If no solution is found, the group can inform the teacher. The appreciation indicators are used by A-EVAL to evaluate the educational session. The teacher uses this information to correct his/her pedagogical scenario.

V. CONCLUSION AND PERSPECTIVES

We have designed an Agent Based System generated by MACADDAM during the generation phase.

All agents of this assistance system have not been yet fully developed. Several issues remain to be explored and implemented. We especially want to make MACADDAM more autonomous and proactive. Thus, the environment should be able to alert the teacher and the students when a group presents a bursting risk or is in a position of educational failure.

The proposed system was coupled to the Learning Management System (LMS) ILIAS and can be adapted to other LMS such as Moodle. However, the number of defined indicators in the system is not very important. We intend to define an indicator library and let each tutor select the indicators that interest him.

Several tools were developed recently for the agent-oriented programming, such as JADE [24], Zeus [25], MadKit [26], Agent Builder [27]. We deploy the system on the multi-agents platform Madkit because MadKit is intended for the development and the execution of multi-agents systems and more particularly for multi-agents systems based on organizational criteria.

REFERENCES

[1] S. Downes, “E-learning 2.0”, eLearn Magazine, http://elearnmag.acm.org/featured.cfm?aid=1104968, October 2005, [retrieve: 05, 2013] .

[2] B. J. Williams and J. Jacobs, “Exploring the use of blogs as learning spaces in the higher education sector”, Australasian Journal of Educational Technology (AJET), Vol. 20, n°2, pp. 232-247, 2004.

[3] P. A. Caron, “Web services plug-in to implement "Dispositives" on Web 2.0 applications”, Proc. Second European Conference on Technology Enhanced Learning (EC-TEL 07), Springer LNCS Crete, Greece, pp. 457-462, 2007.

[4] B. Talon and D. Leclet, “Towards a computer aided pedagogical engineering”, Proc. 3rd International Conference on Computer Supported Education (CSEDU 11), Noordwijkerhout. Netherlands. pp 159-164, 6-8 may, 2011.

[5] D. Leclet, “Environnements Interactifs d’Apprentissage dans des contextes professionnels. Des Tuteurs Intelligents aux Systèmes Supports d’Apprentissage à Distance“, HDR Thesis, Amiens, University Picardie Jules Verne, 2004.

[6] D. Leclet and B. Talon, “La méthode pédagogique MAETIC ». In Libro Veritas (Eds), 2008 .

[7] L. Settouti, N. Guin, A. Mille and V. Luengo, “A Trace-Based Learner Modelling Framework for Technology-Enhanced Learning Systems”, Proc. 10th IEEE International Conference on Advanced Learning Technologies (ICALT 10). Sousse, Tunisia. pp. 73-77, 2010.

[8] D. Bouhineau, J.F. Nicaud, H. Chaachoua, M. Bittar, M. and A. Bronner, “Two Years Of Use Of The Aplusix System”,. 8th IFIP World Conference on Computer in Education, Cape Town, South Africa 2005.

[9] A. Jaillet., « Peut-on repérer les effets de l'apprentissage collaboratif à distance? », Distances et savoirs, Vol. 3, pp. 49-66. DOI : 10.3166/ds.3.49-66 (2005)

[10] O. C. Santos, A. Rodríguez, E. Gaudioso and J.G. Boticario, “Helping the tutor to manage a collaborative task in a web-based learning environment.”, AIED2003 Supplementary Proceedings, Vol. 4, pp. 153-162, 2003.

[11] S. Georges., “Apprentissage collectif à distance, SPLACH: un environnement informatique support d'une pédagogie de projet”. Doctoral dissertation, Université du Maine, 2001

[12] M. May, S. George and P Prévôt, “TrAVis to Enhance Students’ Self-monitoring in Online Learning Supported by Computer-Mediated Communication Tools”,. International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), 3, pp. 623-634, 2011.

[13] J. Ferber, 1999, “Multi-Agent Systems. An Introduction to Distributed Artificial Intelligence”, Vol. 1, Reading: Addison Wesley, London, 1999.

[14] D. Leclet and B. Talon, “MAUI Experiment:a Method for Designing E-Learning Environments in Project Management Training”, Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, ED-MEDIA, Vienna, Austria, AACE/ Springer-Verlag (Ed.), pp. 1-8, 2008.

[15] S. Fiedler, “Personal Webpublishing practices and conversational learning”, Symposium on introducing disruptive technologies for learning: Personal, Webpublishing and Weblogs. ED-MEDIA, Lugano, 2004.

[16] I. Serguievskaia and H. Al-Sakran, “Framework Architecture of e-Loan Negotiation System”, Information and Communication Technologies: From Theory to Applications (ICTTA), pp. 1-6, 2008.

 ISBN: 978-1-61208-280-6

[17] R. Guizzardi-Silva, LM. Aroyo, G. Wagner, "Help&Learn: A peer-to-peer architecture to support knowledge management in collaborative learning communities". Revista Brasileira de Informatica na Educac¸ao, 12 (1). pp. 29-36, 2004.

[18] G. Wagner : "The Agent-Object-Relationship Meta-Model: Towards a Unified View of State and Behavior". Information Systems 28:5, pp. 475504, 2003.

[19] A.J. Fougères, P. Canalda : "iPédagogique : un environnement int´egrant la gestion assist´ee de projets d'´etudiants". Colloque TICE 2002, Lyon.

[20] D. Simian, C. Simian, I. Moisil, I. Pah : Computer Mediated Communication and Collaboration in a Virtual Learning Environment Based on a Multi-agent System With Wasp-Like Behavior in Large-Scale Scientific Computing book 2008, pages 618-625.

[21] A. Mbala, C. Reffay, T. Chanier : "SIGFAD : un système multi-agents pour soutenir les utilisateurs en formation `a distance", 2003.

[22] G. Dyke, "A model for managing and capitalizing on the analyses of traces of activity in collaborative interaction", PhD dissertation, French, 2009.

[23] K. Lund and A. Mille, "Traces, traces d'interactions, traces d'apprentissages : définitions, modèles informatiques, structurations, traitements et usages", (Eds. J.C. Marty & A. Mille). Analyse de traces et Personnalisation des EIAH dans la collection Traité Informatique et Systèmes d'Information *(*dir. *J-C* Pomerol & J-M Labat). Lavoisier-Hermes : Paris, pp. 21-56, 2009.

[24] G. Rimassa, F. Bellifemine and A. Poggi, « JADE - A FIPA Compliant Agent Framework", Proc PMAA 99, pp. 97-108, London, April 1999.

[25] L. C. Lee, D. T. Ndumu and H. S. Nwana, 1998. ZEUS: An Advanced Tool-Kit for Engineering Distributed Multi-Agent Systems, In Proceedings of the Practical Application of Intelligent Agents and Multi-Agent Systems (PAAM 98), pp. 377-392, London, 1998.

[26] O. Gutknecht and J. Ferber, 2000. Madkit: a Generic Multi-Agent Platform, Proc. 4th International conference on Autonomous Agents (AGENTS 2000), Barcelona, ACM Press, pp. 78-79, 2000.

[27] AgentBuilder U.G., An Integrated Toolkit for Constructing Intelligent Software Agents, AgentBuilder, User's Guide, April 2000

 ISBN: 978-1-61208-280-6

A Multi-Agent System to Implement a Collaborative Learning Method

Hanaa Mazyad, Insaf Tnazefti-Kerkeni and Henri Basson
Laboratoire d'Informatique, Signal et Image de la Côte d'Opale
Université de Lille Nord de France
Calais, France
e-mail: {mazyad, kerkeni,basson}@lisic.univ-littoral.fr

Abstract— **This paper presents a multi-agent system with Peer-to-Peer architecture that implements a pedagogic method called MAETIC. Indeed, the system objective is to allow the collaborative learning of project management. However, collaborative e-learning imply new roles for tutors as well as for learners. It is therefore essential to identify the users' needs and integrate, in the system, the functionalities which allow satisfying such needs. In addition, it is essential to avoid the failure and/or desertion of learners by providing tutors and learners with the opportunity to obtain information about the progress of their learning processes as well as the level of collaboration and sociability of each learner in the group**

Keywords- Collaborative learning; Multi-agent systems; Peer-to-Peer systems

I. INTRODUCTION

The Information and Communication Technologies in Education (ICTE) allowed users dispersed geographically to work and thus maximize the creativity and efficiency of group learning. However, the effectiveness and efficiency of collaborative learning depend on the motivation of its members to collaborate, on the number of members, on the time they can consecrate to this work and on their skills. Thus, the responsibility becomes collective.

In this paper, we are interested in the field of collaborative learning and especially on a teaching method called MAETIC (from french « Méthode pédAgogique InstrumentEe par les TIC), equipped with ICT (Information and Communication Technologies), aiming at developing a project-based learning pedagogy [1]. The main purpose of the deployed pedagogy is to provide real support to the management of project-based learning by group of students. For this, a learning platform has been established to offer students a support for the development of their projects and monitoring by the teacher. However, the services offered by the system suffer from a lack of support tools for managing the archiving of interactions in the system. Indeed, in the field of collaborative learning, management of traces is important since it allows analyzing the collected information concerning the learner or group of learners and provides the tutor with accurate and adequate information for his needs on the individual and collective evolution of learning.

In this paper, we describe a multi-agent system (MAS) for the collaborative learning of project management. This system implements the MAETIC method and provides users with functionalities to keep a detailed history of all the actions of the groups when they access the system in order to assess the group's life and its evolution.

This paper is organized in 6 sections. In Section II, the MAETIC method is briefly introduced. Some related work is presented in Section III. Then, in Section IV, an agent-oriented collaborative learning system that implements MAETIC is proposed. In Section V, some obtained results are presented. Finally, conclusion and some perspectives are given in Section VI.

II. MAETIC METHOD

This section describes the MAETIC method and presents the needs to be taken into account in order to implement this method. In addition, it defends the use of multi-agents and Peer-to-Peer (P2P) systems by presenting their contribution.

A. Description of the method

MAETIC is a pedagogic method instrumented by the ICT (Information and Communication Technologies). It is a teaching method which, as part of pedagogy for project-based learning, describes a set of formalized and applied procedures according to defined principles. Thus, the objective of MAETIC is to allow students to develop requested knowledge and skills. For the teachers, MAETIC's objective is to promote the establishment of a process that facilitates their educational activities.

MAETIC is based on five stages commonly adopted in the process of project management [2]: the initialization, the preparation, the planning, the project monitoring and the revenue (Figure 1). Each stage establishes activities, requires the production of one or more deliverables, and takes place over one or several sessions. Since the work is collective, MAETIC advocates the establishment of an organization in the group project (description of roles) that promotes the acquisition or strengthening of transversal skills needed for teamwork. The fact of making the group produce deliverable develops qualities related to the written production.

Thus, each project team must establish its weblog. This weblog aims to describe the life of the project. Besides the general information on the project (subject, members), it is responsible of storing all the notes concerning the project's life and is also responsible for collecting developed deliverables.

 ISBN: 978-1-61208-280-6

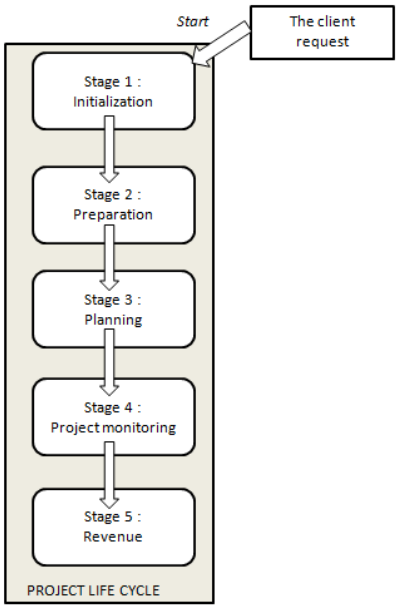


Figure 1. The five stages of MAETIC method

B. *Platform needs and contribution of multi-agents system*

MAETIC, as we have seen, is primarily concerned with implementing a project-based learning pedagogy to promote the learning of knowledge and skills that are essential to exercise a profession. Communication between the different actors: the director, teachers and learners in the initial platform is via the Weblog. Weblog technology was chosen for its maneuverability. The establishment phase of a weblog is very short and its use is accessible to all. In addition, weblogs can create a social bond with the students and seems to facilitate the writing of learners through the "posts" [3]. However, this technology provides a hardware consequent, very easy to collect but more difficult to analyze. The time spent by the teacher to monitor and analyze the activities of the learner is higher than that spent in the classical classes.

In addition, in the context of e-Learning, it is essential to avoid the failure and/or desertion of learners, to propose for tutors automatic tools, with assistance and decision support. These tools help to keep traces of all interactions related to students belonging to a given group and report the indicators of progression status in this learners group and its durability.

In a collaborative learning system, each member must manage and exchange his knowledge and cooperate with others in order to achieve his goals. Compared to these aspects, a Peer-to-peer system is particularly suitable to develop a collaborative learning system due to the following capabilities:

- It supports autonomy: every member of the system is seen as a peer that manages and has control over a set of local technologies, applications and services;
- It is decentralized: the community of peers is able to achieve its goal independently from any specific member or component;
- It is cooperative: in order to join and use the system, each member must provide resources or services to the others;
- It is dynamic: peers and resources can be added or removed at any time.

The multi-agent system is also an appropriate framework for realizing a P2P applications. The characteristics that they have, especially (a) their capability to allow the sharing or distribution of knowledge, and (b) that they assemble a set of agents and coordinate their actions in an environment to accomplish a common goal, are needed in this P2P application.

III. RELATED WORK

Several works on collaborative e-learning exist in the literature. Formid [4] for example, is an object-oriented platform with client-server architecture. This platform provides teachers with information about the realization of students' activities. However, we are interested by agent-oriented work. So, we review the approaches presented in the literature for providing multi-agents platforms for collaborative learning.

In the field of Artificial Intelligence and education, several approaches have been developed. For example, Guizzardi et al. [5] have developed a Peer-To-Peer system called "Help & Learn". This system was modeled using an agent-oriented language called AORML [6]. It is an open, centralized system that is designed to support the extra-class interactions between students and teachers. "Help & Learn" is limited to providing assistance to students who request it. Other systems have been developed. Fougeres and Ospina [7] have proposed a based-agent mediation system for the project management platform called iPédagogique. This system, modeled in AUML, serves as an interface between the human and the application to enhance their relationship and is used to promote collaboration among users. None of these two systems cares about monitoring learning and therefore, cannot trace user activities. Another system was proposed recently called I-MINDS [8]. This system includes a teacher agent for supporting the instructor. The teacher agent allows the instructor to interact with students, manage questions and answers (Q&A) sessions, administer quizzes, post evaluations, form groups and monitor individual and group performances. For computer-supported collaborative learning, I-MINDS provides a student agent for each student. Each student agent monitors and models its user. I-MINDS is interested in monitoring learning. However, it is not sufficiently independent and does not start up alerts to prevent teachers if there is a problem with a student or group. In addition, it does not give information about learners and groups collaborative level.

IV. TOWARDS AN AGENT-ORIENTED COLLABORATIVE LEARNING SYSTEM

The agentification of MAETIC method needs the use of a methodology. Several methodologies were proposed for the

 ISBN: 978-1-61208-280-6

development of MAS. We will follow Aalaadin methodology [9]. Aalaadin supports:

- Distributed systems where their members are geographically distributed.
- Opened systems that allow the integration of new members or the departure of a member.
- Complexity of elaborated systems.
- Autonomy of agents.
- Cooperative behavior of agents to achieve their goals.

A. *The system modeling*

Aalaadin is an organizational method developed by Gutknecht and Ferber. It is, first, a background for developing multi-agent systems, providing methodological guidance and secondly, a prototyping and running environment for agents based on notions of group and role through the AGR (Agent/Group/Role) model.

The AGR model of Aalaadin methodology is based on three primitive concepts, Agent, Group and Role that are structurally connected and cannot be defined by other primitives.

Agent: an agent is an active and communicating entity that plays roles within groups. An agent may hold multiple roles, and may be a member of several groups. An important characteristic of the AGR model is that no constraints are placed upon the architecture of an agent or about its mental capabilities. Thus, an agent may be reactive as an ant, or clever as a human.

Group: a group is a set of agents sharing some common characteristics. A group is used as a context for a pattern of activities, and is used for partitioning organizations. Two agents may communicate if and only if they belong to the same group, but an agent may belong to several groups. This feature will allow the definition of organizational structures.

Roles are local to groups, and a role must be requested by an agent. A role may be played by several agents.

1) The analysis phase : This phase identifies the system functions and dependencies within the identified communities. The functionalities of the system have been organized into the following objectives: a) Provide learners and tutors with an environment for project-based learning with tools allowing collaborative learning; b) Support the learners and help them in the achievement of their project by providing them with functionalities for monitoring their activities and those of the group ; c) Analyze the information collected about the learner or group of learners; d) Provide tutors with accurate and adequate information for their own needs on the individual and collective evolution of learning; e) Determine the level of productivity of each learner in terms of achieving pedagogic activities, as well as his level of communication with other members of the group; f) Determine the progress of each group regarding to the achievement of activities but also to the present, absent, inactive members; g) Determine the levels of activities achievement by all the groups and adjust the calendar if necessary; h) Provide users with information about the use of collaborative tools; i) Determine the collaborative work level in each group; j) Enable learners to also have access to this information which supposed to empower them, show their performance and enable them to compare this performance with other learners in the same group.

2) The design phase : This section presents the system design. Indeed, to design the system, the functionalities listed in the analysis phase will be taken into account to design the system groups' structure. In addition, the requirements of P2P systems will be followed to design databases.

a) Databases Design : P2P systems require from each peer to be client and server at the same time. Thus, there is

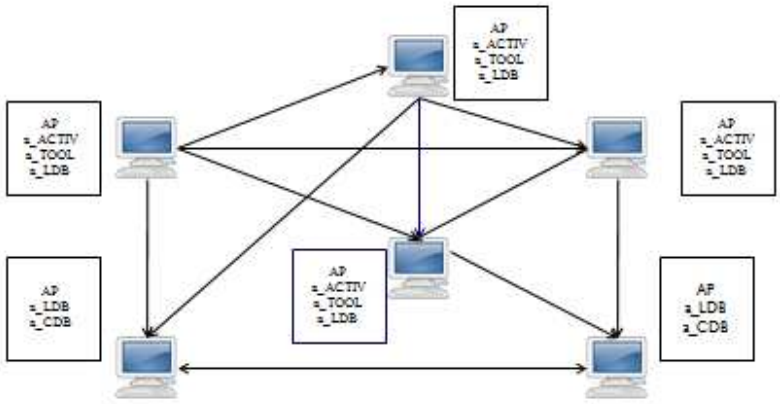


Figure 2. Agents deployment

Role: the role is the abstract representation of a functional position of an agent in a group. An agent must play a role in a group, but an agent may play several roles.

no centralized server which contains all data and files of each user. Thus, in the proposed system, each user will have its local database (L_DB) containing his data. The user must

allow other users of his group to have access to this database in order to recuperate files or data. However, to enable the user to know the list of his group users, a centralized entity is used. Indeed, a hybrid architecture is adopted which means that user machines are connected to a machine called super-peer that plays the role of a server containing a directory. This directory contains the user information (username, IP address, etc.). However, this central database will not contain user files. On the contrary, file and data exchange will be based on the P2P model where user machines are both client and server.

b) Groups Design : There are two group structures in our system: the tutors' group structure and the learners' group structure. The tutors' group structure has the roles played by tutors: a) Guide the learners and adjust the pedagogic scenario; b) Encourage collaborative work; c) Motivate struggling learners. The learners' group contains the roles below:

- Helper: giving help to learners;
- Help consumer: asking questions to other learners;
- Author: upload documents to give it to tutors
- Learner: download documents given by the tutors

Each learner and tutor has a personal assistant agent (AP). The "AP" agent plays roles enumerated above and executes all the request of the user.

To collect and manage traces, we need to define the following roles:

- Supervisor of activity
- Supervisor of tools
- Local database manager (L_DB)
- Central database manager (C_DB).

We attribute each one of these roles to one agent:

- a_ACTIV: supervises users' activities during a session. It provides statistics concerning the progression of each activity. It reminds learners about deadlines and notifies the late groups by sending alerts.
- a_TOOL: supervises the use of tools. It provides statistics about the use of collaborative tools (Email, Chat, Forum, etc.). In fact, the "a_TOOL" agent saves in the "L_DB" of each user, tools that he used;
- a_LDB: manage the interactions between system agents and the "L_DB" and between "L_DB" and "C_DB".
- a_CDB: manages the central database.

Figure 2 shows the agents deployment. The agent "a_CDB" is a centralized entity which exists only in the tutor's space. It should be noted that the agents "a_ACTIV" and "a_TOOL" do not exist in the tutor's space due to the fact that the system does not collect their information.

3) The realization phase : We have chosen to deploy the system on the multi-agents platform Madkit [10] [11]. This choice is taken due to the fact that MadKit is intended for the development and the execution of multi-agents systems and more particularly for multi-agents systems based on organizational criteria (groups and roles). However, MadKit does not impose any particular architecture to the agents. MadKit communication is based on a peer to peer mechanism, and allows developers to quickly develop distributed applications using multi-agents principles. Concerning the database management, Java has the Java

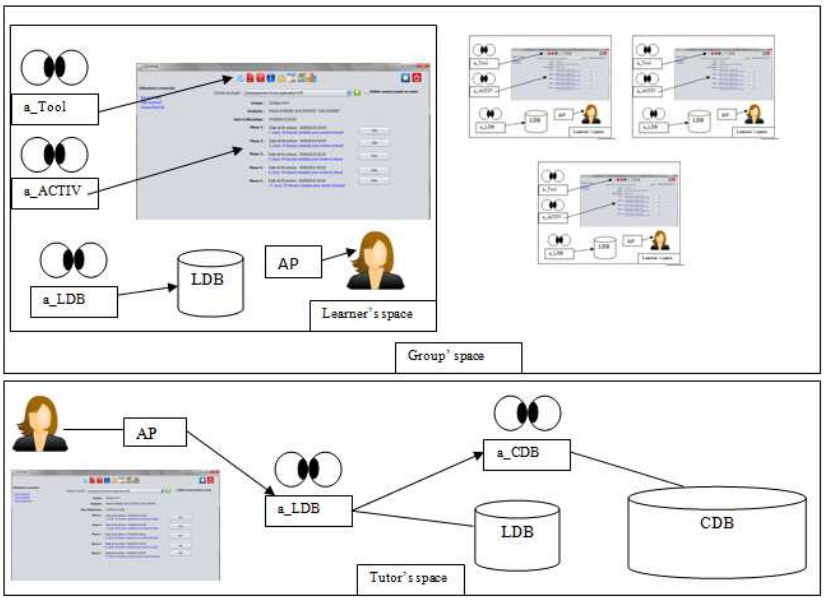


Figure 3. The system agents in their respective space

 ISBN: 978-1-61208-280-6

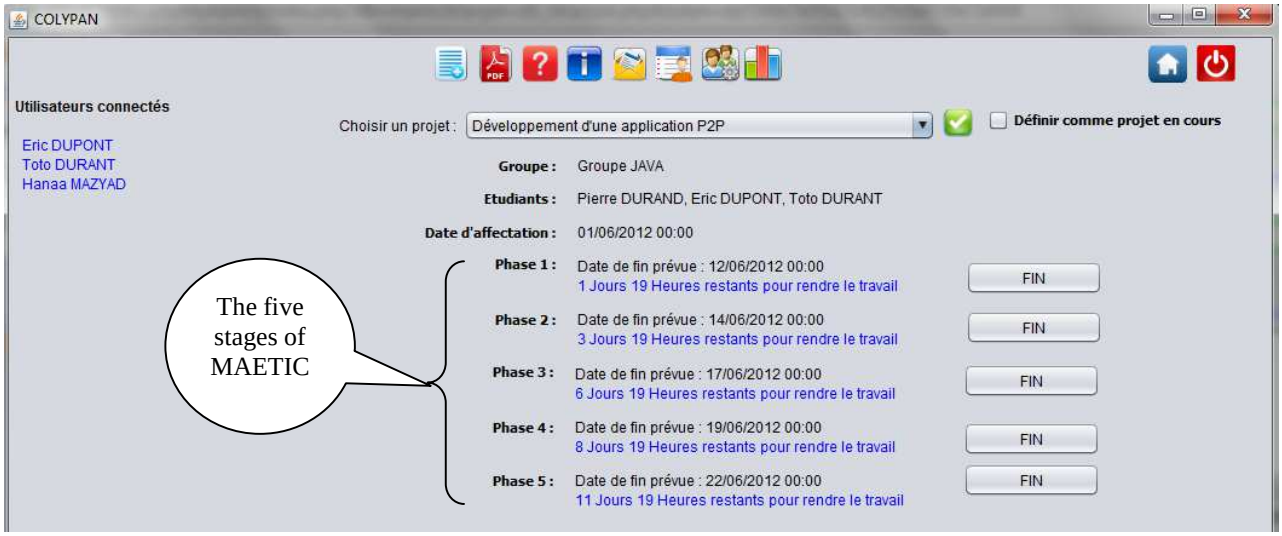


Figure 4. The graphic interface

Database Connectivity (JDBC) API that allows the connection to databases and is independent of database management system (DBMS). MySQL is chosen as DBMS because it is a popular server used for free. It runs on most operating systems and is often used in conjunction with Java.

As the hybrid model of P2P requires, our system consists of peers connected to a super-peer. Thus, machines with low bandwidth are called peers. Machines with good bandwidth are super-peers. Super-peers play the role of localization server. This model has several advantages

- It decreases of the number of connections on each server which helps in avoiding bandwidth problems.
- It uses a mechanism of P2P decentralized model to update a client directory and files indexes from information coming from other servers. Thus, a server can provide any client with all the information on the network.
- It allows identifying the system users which is essential in the learning context, while keeping the advantages of P2P systems.

In the proposed system, the system's central database is installed on the super-peer. However, at no time user files will be stored on the server. Indeed, the server can localize these files and users can get them from the machine where they are stored in a shared folder and accessible remotely. In addition to this central database, each user has on his machine a local database that contains information about him and the data he owns. Users can access this local database as long as the owner allows them.

We created our agents using the Madkit agent « Designer ». It is a tool designed to facilitate the agent creation and launch. However, this agent does not require any running concept. In fact, it only provides methods to start the life cycle of an agent (Activate, live, end) and the user must program its behavior.

Figure 3 shows the agents interacting in users (Tutor, Learner) space. In each learner's space, "a_ACTIV" agent collects information about the learner activities and "a_TOOL" agent collects information about the use of tools in this space. Then, these two agents send the collected information to the "a_LDB" agent. Finally, "a_LDB" agent sends statistics based on this collected information to the "a_CDB" agent. In the system, these statistics will be sent to the "a_CDB" agent each 15 minutes as long as the learner is connected.

V. OBTAINED SYSTEM

Our system is a collaborative learning system that implements the MAETIC method. It aims to enable collaborative learning of project management. In this system, the tutor's role is to guide learners and facilitate their learning. The group's role is to motivate learners and create a social link that prevents the feeling of isolation. Learners learn by interacting with others. In addition, agents interact and communicate to achieve a common goal: satisfy the users (learner or tutor) requirements. Indeed, this system provides tutors and learners with information on group progress and on collaboration and sociability levels of each learner, and allows each learner to perceive his own learning situation. In addition, it is independent and proactive. It scrutinizes periodically the data of interaction and triggers alerts to prevent the tutor and the learner when the group is at risk of bursting or is experiencing educational failure.

Figure 4 shows the system's graphic interface. Firstly, learner should be a member of at least one group to be able to work on a project. Secondly, learner chooses the project that he wants to work on and then, he can start his learning process and system's agents start collecting information. In

 ISBN: 978-1-61208-280-6

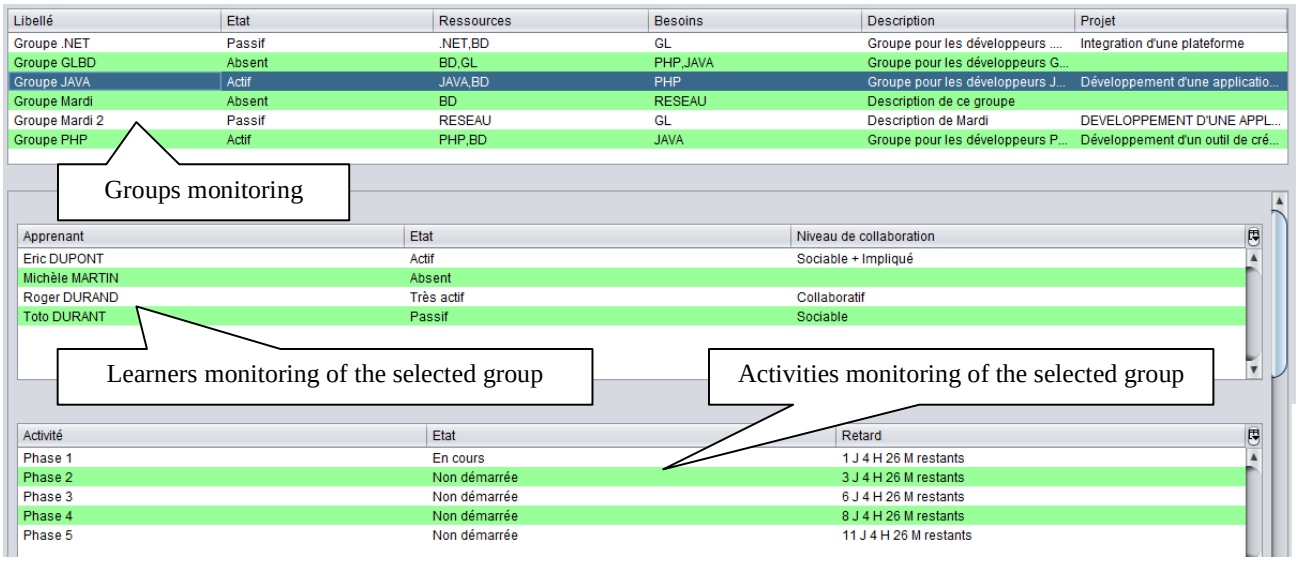


Figure 5. The group's monitoring

the system, learner can be member of several groups but in each group, he works on a single and different project.

Figure 5 shows the group monitoring. It presents information about each member of the group but also about the realization of group's activities.

VI. CONCLUSION

In this paper, we are interested in the design and the realization of a system that implements the MAETIC method. We have chosen to use multi-agents systems with P2P architecture to realize this collaborative e-learning system. In this system, each learner contributes in the learning process of the group, and in return, the group contributes in the learning process of its members. The consistency of the whole group allows achieving the goal. However, collaborative e-learning implies new roles for tutors as well as for learners. Thus, the user's needs are identified and functionalities which allow satisfying such needs are integrated into the system. In addition, this system provides tutors and learners with the opportunity to obtain information about the progress of their learning processes as well as the level of collaboration and sociability of each learner in the group.

At present, several points need to be explored. Without being exhaustive, the e-learning is a multidisciplinary field ranging from the social sciences to the hard sciences, in which users with heterogeneous profile interact. We think it will be interesting in the field of collaborative learning to propose analysis methods of chat and forums content.

REFERENCES

[1] D. Leclet and B. Talon, "Assessment of a Method for Designing E-Learning Devices". Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, ED-MEDIA, Vienna, Austria, AACE/ Springer-Verlag (Ed.), 2008, pp 1-8.

[2] H. Marchat, Kit de conduite de projet. Paris: Organization editions, 2001.

[3] S. Fiedler, "Personal Web publishing practices and conversational learning". Symposium on Introducing disruptive technologies for learning:. ED-MEDIA Lugano, 2004, pp. 2584-2591.

[4] V. Guéraud and J-M. Cagnat, "Automatic Semantic Activity Monitoring of Distance learners Guided by Pedagogical Scenarios", Innovative Approaches for Learning and Knowledge Sharing, EC-TEL 06, Lecture Notes in Computer Science LNCS 4227, Springer 2006, pp.

[5] R. Guizzardi-Silva, LM. Aroyo and G. Wagner, "Help&Learn: A peer-to-peer architecture to support knowledge management in collaborative learning communities". Revista Brasileira de Informatica na Educação, 12 (1),2004, pp. 29-36.

[6] G. Wagner, "The Agent-Object-Relationship Meta-Model: Towards a Unified View of State and Behavior". Information Systems 28:5, 2003, pp. 475–504.

[7] A.J. Fougères, "Vers un Système de médiation pour les systèmes coopératifs ". Mémoire d'habilitation à diriger des recherches, Université de Technologie de Belfort-Montbéliard, 2010, pp. 76-78.

[8] N. Khandaker, L. -K. Kiat, L. D. Miller and H. Jiang, "Lessons Learned from comprehensive deployments of multi-agent CSCL applications I-MINDS and classroom wiki". IEEE Transactions on Learning Technologies, Volume 4, Number 1 (TLT4(1)), January 2011, pp. 47-58.

[9] J. Ferber and O. Gutknecht, "Aalaadin: a meta-model for the analysis and design of organizations in multi-agent systems ". In : Demazeau, Y., éditeur : 3rd International Conference on Multi-Agent Systems, Paris, IEEE, 1998, pp. 128–135.

[10] O. Gutknecht and J. Ferber, " Madkit : a generic multi-agent platform". Proceedings of the fourth international conference on autonomous agents, 2000, pp. 78-79.

[11] www.madkit.org

 ISBN: 978-1-61208-280-6